The role of subgroups and sub-populations in drug development and drug regulation

Thesis

THE ROLE OF SUBGROUPS AND SUB-POPULATIONS IN DRUG DEVELOPMENT AND DRUG REGULATION

Andrew Garrett, BSc. MSc. CStat

Date: 6th April 2006

Research thesis submitted for the award of PhD

Department of Statistics

Faculty of Mathematics and Computing

The Open University

THE ROLE OF SUBGROUPS AND SUB-POPULATIONS IN DRUG DEVELOPMENT AND DRUG REGULATION

Andrew Garrett, BSc, MSc, CStat

THESIS ABSTRACT

This thesis addresses the role of subgroups and sub-populations in drug development and regulation and includes the critical appraisal of regulatory guidance.

Chapter One introduces clinical trial methodology and describes the current regulatory environment.

In Chapter Two, randomisation is reviewed in relation to unbiased estimation of treatment differences and the impact of data exclusion to form subsets is described.

Simpson's paradox (SP) is considered in Chapter Three. Randomisation is shown to protect against SP, while a treatment by factor interaction is not required. Balance is re-defined for the odds formulation leading to identical unconditional and conditional parameters. The chances of SP (and less extreme inconsistencies) occurring are quantified using simulation, with varying sample size.

Chapter Four considers treatment by subgroup interactions. Suggestions regarding the magnitude of a clinically relevant interaction are presented while a simple Bayesian approach to evaluate interactions using margins for the interaction parameter is applied to published data.

Chapter Five considers non-inferiority in relation to sub-populations and covariate adjustment for binary outcomes. It is shown that the Per Protocol population is not necessarily conservative and simulation is used to demonstrate the impact on the type I and II errors. Using simulations it is shown that an increase in the type I error occurs if an important covariate is excluded from the logistic model when testing for non-inferiority.

Chapter Six is directed towards the sub-population of children. The impact of off-label treatment is discussed in relation to the ethics of placebo-controlled trials, together with the importance of randomisation in evaluating long-term safety.

In Chapter Seven, a therapeutic area is selected to illustrate the challenges raised during the previous chapters and wording changes to current regulatory guidelines are proposed. The thesis closes with personal thoughts regarding the future potential for individualised treatment.

TABLE OF CONTENTS

ACKNOWLEGEMENTS

In conducting this research, I would like to thank my external supervisor, Professor Stephen Senn, for his support, encouragement and comment during the seven years of study. I would also like to acknowledge the positive impact of many of his thought provoking papers that stimulated my interest in many of the areas covered in this research thesis. I would also like to thank Professor David Hand for agreeing to take on a part-time mature student in the first place all those years ago and for providing valuable comment on my work.

During his time at Quintiles, Dr Dennis Chanter was a valued sounding board and in particular provided careful reading and valuable comment on my paper entitled *Therapeutic equivalence: fallacies and falsification.*

From a purely practical point of view, I would like to highlight my great debt to Isabel Bentley at Quintiles Information Services in Edinburgh who has never failed to track down a paper - no matter how obscure - within a timeframe that still amazes me. I would also like to acknowledge Quintiles Limited for providing financial support in terms of research fees.

Finally I would like to acknowledge my young children, Christian and Paddy, who on occasions had to endure a grumpy father who was in search of peace and quiet, and my wife, Emma, who tolerated us all and just about kept us sane. To them, I dedicate this research thesis.

ABBREVIATIONS

6MP	6-mercaptopurine
AAP	American Academy of Pediatrics
ABPI	Association of the British Pharmaceutical Industry
ADHD	Attention Deficit Hypersensitivity Disorder
AIDS	Acquired Immune Deficiency Syndrome
ANCOVA	Analysis of Covariance
BMJ	British Medical Journal
BUGS	Bayesian inference Using Gibbs Sampling
CDER	Center for Drug Evaluation and Research
CFC	Chlorofluorocarbons
CHMP	Committee for Medicinal Products for Human Use
CI	Confidence Interval
CNS	Central Nervous System
CPMP	Committee for Proprietary Medicinal Products
CTA	Clinical Trial Authorisation
CTD	Common Technical Document
CR	Complete Response
DBP	Diastolic Blood Pressure
DSMB	Data Safety Monitoring Board
DH	Declaration of Helsinki
DHSS	Department of Health and Human Services
EMEA	European Agency for the Evaluation of Medicinal Products
EU	European Union
F	Factor
FDA	Food and Drug Administration
FDAMA	Food and Drug Administration Modernization Act
FS	Full Set
GACVS	Global Advisory Committee on Vaccine Safety
GAO	General Accounting Office
HIV	Human Immunodeficiency Virus
ICH	International Conference on Harmonisation
IND	Investigational New Drug
IHS	Immunocompromised Host Society
ISE	Integrated Summary of Efficacy
ISS	Integrated Summary of Safety
ITT	Intent-to-treat
JAMA	Journal of the American Medical Association
LCL	Lower Confidence Limit
LOCF	Last Observation Carried Forward
LR	Likelihood Ratio
MCA	Medicines Control Agency
MCMC	Markov Chain Monte Carlo
MHRA	Medicines and Healthcare products Regulatory Agency
MMR	Measles, Mumps, Rubella vaccine
NDA	New Drug Application
NEJM	New England Journal of Medicine
NIH	National Institute of Health

NNT	Number Needed to Treat
NSAID	Non-steroidal Anti-inflammatory Drug
OR	Odds Ratio
OS	Open Surgery
PD	Pharmacodynamic
PDT	Photodynamic Therapy
PK	Pharmacokinetic
PN	Percutaneous nephrolithotomy
PP	Per Protocol
PPHN	Persistent Pulmonary Hypertension
PR	Partial Response
PSI	Statisticians in the Pharmaceutical Industry
PSUR	Periodic Safety Update Reports
PtC	Points to Consider
PV	Protocol Violator
QLT	Qualitative
QNT	Quantitative
RCT	Randomised and Controlled Clinical Trial
RSV	Respiratory Syncytial Virus
SACHRP	Secretary's Advisory Committee on Human Research Protection
SE	Standard Error
SP	Simpson's Paradox
SR	Standardised Range
SS	Sum of Squares
SUSARS	Suspected Unexpected Serious Adverse Reactions
T	Treatment
UCL	Upper Confidence Limit
UK	United Kingdom
US	United States of America
WHO	World Health Organisation

LIST OF TABLES

LIST OF FIGURES

PUBLICATIONS

Garrett AD. Therapeutic equivalence: fallacies and falsification. *Statistics in Medicine* 2003; **22**: 741-762.

RELATED PRESENTATIONS

Rare - but Well Done. Developing Drugs for Rare and Life-threatening Diseases (2004). DIA Conference, Washington, USA.

Tutorial Case Studies: Stratified Binary Data and Logistic Regression (2004). PSI Conference, Chester.

Switching between Superiority and Non-inferiority - Ensuring a Coherent Approach (2003).
Conference of the Japanese Society for Computational Statistics, Gotemba City, Japan.

The Evaluation and Interpretation of Treatment by Subgroup Interactions (2003). PSI Conference, Bristol.

Switching between Superiority and Non-inferiority - Ensuring a Coherent Approach (2003). Briefing on the latest CPMP Points to Consider documents.
Henry Stewart Conference. London.

Therapeutic Equivalence: Problems and Practical Solutions (2002). Statistical Techniques used in Clinical Trials.
Henry Stewart Conference. London.

Therapeutic Equivalence: Fallacies and Falsification (2001)
ISCB Conference, Stockholm, Sweden.

Simpson's Paradox and Related Inconsistencies in Drug Development and Regulation (1998).
PSI Conference, Harrogate.

Issues Arising from Subgroup Analyses (1997). Understanding, Applying and Not Misusing the Mathematical and Statistical Techniques Used in Clinical Trials.
Henry Stewart Conference Studies. London.

PREFACE

We'll drink a drink a drink
To Lily the Pink the Pink the Pink
The saviour of the human race
For she invented medicinal compound
Most efficacious in every case

Gorman, M^{c}Gear and M^{c}Gough, 1968
The Scaffold, Parlophone

Around the world, the development of new pharmaceutical treatments is subject to strict regulatory control, and the accretion of evidence to support their approval for the treatment of patients takes many years. The culmination of research effort, costing in the region of $800m (Tufts CDD, 2001), is the creation of a so-called Common Technical Document (CTD), which is used by pharmaceutical companies to summarise their evidence (ICH M4, 2000).

Regulatory authorities regard robustness as an important property of the evidence presented in the CTD. The adjective robust means *strong, uncompromising* and *vigorous*, and is derived from the Latin word *robustus* meaning *oaken, solid, firm and hard.* The Oxford dictionary (1993) includes a description of robust that captures the essence of the pharmaceutical companies' challenge when presenting their evidence: *designating a result where the result is largely independent of certain aspects of the input.* This description highlights the dual requirement to study a new treatment under a broad range of conditions, and furthermore to evaluate the data generated for consistency of effect under these various conditions – for instance, across sub-populations and subgroups. As such, the generation of robust evidence does not simply equate to the accumulation of ever increasing amounts

of similar data. Indeed Hempel (1966) wrote that *the confirmation of a hypothesis depends not only on the quantity of the favourable evidence available, but also on the variety: the greater the variety, the stronger the resulting support.* Variety is therefore a key component in the generation of robust evidence and to the development of safe and effective treatments.

Now, variety can be introduced and carefully controlled at the design stage of a study. Drug developers set the range of inclusion and exclusion criteria for patients and select the countries and study centres to conduct the investigations. The duration, dose and frequency of drug administration are chosen and also the type of control treatment. Furthermore the study procedures will be carefully detailed in the study protocol. However the investigation of variety occurs once the data are collected – that is, at the analysis stage.

In drug development the ideal scenario for a new treatment is that it is uniformly safe and efficacious in a particular disease area. Unfortunately, this is rarely the case and regulatory authorities need to be made aware of the circumstances under which modification or restriction of use is required. (For example, the analgesic effect of codeine is almost absent in around 7% of Caucasians (FDA, 2002) due to the lack of a specific liver enzyme which is required to convert codeine to its active metabolite, morphine.) Such modifications and restrictions are discovered through the estimation of treatment effects through various slices of the data contained within the CTD. However the regulatory authorities also acknowledge the limitations of such an approach - especially with regard to safety. Since a typical CTD would include in the region of 1500 patients treated with the new drug (FDA, 2002) – many of whom will have had only limited exposure – the detection and attribution of rare adverse drug reactions is unlikely therefore, regardless of

the degree of variety present. As a consequence, post approval monitoring is recognised as being important in the ongoing evaluation of safety. Nevertheless, despite the limitations of the safety evaluation due to its multidimensional nature, the evaluation of efficacy is often more informative.

This research thesis is directed towards an investigation of some of the issues surrounding the generation of robust evidence to support the review and approval of new pharmaceutical treatments. Specifically, I will examine two related areas where the evaluation of consistency of effect has been of considerable regulatory interest – that is, the choice of patient analysis populations, and the investigation of patient subgroups. Accordingly, regulatory considerations will be a running theme throughout and, where applicable, current regulatory guidelines will be critically appraised in relation to my findings.

In Chapter 1, I begin by providing the reader with the necessary background information to place this research effort in context. Firstly, pharmaceutical research – that is, drug discovery and drug development - will be described to provide the reader with an understanding of the strictly sequential nature of the investigations and how knowledge is acquired and built upon through time. Secondly, the regulatory environment that has led to the evolution of this linear development process will be described. This section will include a brief history of drug regulation, explain the current regulatory structures and processes, and describe how attitudes have evolved in this area. Furthermore the specific guidelines that increasingly drive the design and analysis of clinical trials will be discussed. Thirdly, the randomised and controlled clinical trial (RCT) will be described. The RCT underpins the confirmatory phase of drug development and an understanding of

the basic principles of this trial design is key to explaining the roles and importance of patient sub-populations and subgroups in the evaluation of robustness.

A review of the basic principles surrounding the sub-setting of clinical trial data is provided in Chapter 2. This will focus on randomisation as the basis for providing unbiased estimates of treatment effects, and explain how the exclusion of data to form sub-populations and subgroups has the potential to introduce bias into the estimation process. The so-called intent-to-treat principle will be reviewed and consideration given to its evolution, perceived conservative nature and its practical implementation. The review will then proceed to consider patient subgroups and the methods employed to investigate treatment effects within subgroups and the consistency of effect between subgroups. In particular, multiplicity and the risk of observing inconsistent results from subgroup to subgroup will be examined.

An investigation of Simpson's paradox (SP) in the clinical trial setting is the basis of the ideas developed in Chapter 3. SP describes a reversal effect whereby the differences between treatments from all subgroups are in the opposite direction to the overall difference between treatments. This phenomenon is well documented in the statistical literature, although it will be shown that SP does not require the presence of a treatment by subgroup interaction as stated previously (Nelder, 1994b). It will be shown how treatment estimates are protected against SP through the use of randomisation and stratification, while the chances of observing SP in the clinical trial setting will be quantified as being small through a series of simulation exercises. SP will be used to provide a general mechanism for observing inconsistent results including cases where the overall treatment difference is greater than that observed in all subgroups or *vice versa*. Although both

binary and normally distributed data will be investigated, focus will centre on the odds model formulation for binary data since it behaves in an unusual manner.

Chapter 4 is a natural progression from Chapter 3, since it investigates potential interactions between treatments and subgroups in the clinical trial setting. Initially the relationship between the overall weighted treatment difference and the subgroup treatment differences will be examined and the correlation structure determined. An attempt will then be made to present a unified approach to the evaluation of interactions. In this respect the approach taken to distinguish interactions of a qualitative or quantitative nature will be compared to the approach of estimating the magnitude of the interaction parameter. This will lead to suggestions for determining the magnitude of a clinically relevant interaction effect. It will then be shown that for the primary endpoint, a superiority trial will *a priori* have similar power to detect the pre-specified treatment difference, as it will have to detect a qualitative interaction. Finally, a Bayesian approach to evaluate and interpret interactions, using margins for the interaction parameter (along similar lines to equivalence studies), will be presented and applied to some published clinical trial data.

Chapter 5 switches tack and focuses upon a current controversial topic in the clinical trial methodology - the issues of therapeutic equivalence and non-inferiority. This chapter investigates how the areas of sub-populations and subgroups impact the evaluation of equivalence and non-inferiority. After describing the methodological, philosophical and regulatory background of equivalence and non-inferiority, some initial work is undertaken to highlight the problems with current methods of margin specification. In particular, there will considerable focus on the odds ratio for binary outcome data. Sub-populations will then be considered and it will be shown that the Per Protocol (PP) population is not necessarily conservative from a regulatory perspective, as some have stated, while the

impact of using the PP population in comparison with a the full population will be quantified in a series of simulations. Finally subgroups will be considered in terms of the impact on the estimate of the treatment difference of including a two-level prognostic factor in the logistic regression model. A series of simulations will be used to demonstrate modest inflation of the type I error rate, when testing for non-inferiority, in cases where an influential factor is excluded from the model. Consistency issues will also be highlighted, in relation to sub-populations and covariate adjustment, when switching hypotheses in the same study from non-inferiority to superiority or *vice versa.*

In Chapter 6, I have selected a specific patient population for investigation that has been described in the past as *the therapeutic orphan* (Shirkey, 1968). Children represent around one half of the world's population of six billion persons but drug labelling frequently discourages the use of drugs in this paediatric population. This is not necessarily because the drugs are unsafe, but rather as an indirect consequence of inadequate clinical trials research in this area. This situation is now being rectified and clinical trials in the paediatric population is now an integral part of the drug development process. However these trials provide a unique set of challenges and in this chapter I will highlight some of these, including the design of studies where unapproved treatments are regarded as standard care, and the investigation of the long-term effect of treatment on child development.

Chapter 7 brings together the findings presented in earlier chapters and discusses their relevance to the generation of robust evidence for new pharmaceutical treatments. In this respect, I have selected a specific therapeutic area where some practical solutions will be proposed to the specific challenges raised. Modifications to the text of current regulatory guidance will also be proposed. Consideration will then be given to the potential use of

genetic information to tailor treatments to individual patients. In particular, I will caution against the over optimistic belief that that we will move from a stochastic to a deterministic approach to drug treatment. Finally I will consider the generalisation of clinical trial outcomes to the future treatment of patients and question the future direction of drug development.

CHAPTER ONE: BACKGROUND

Mr Frears
Had sticky-out ears
And it made him awful shy
And so they gave him medicinal compound
And now he's learning how to fly.

1.1 INTRODUCTION

The intention of this introductory chapter is to provide the background information necessary to place this research effort in context. When considering sub-populations and subgroups in clinical trial research, it is important to understand the building blocks of the drug development process and to appreciate the prevailing climate of drug regulation. Accordingly, pharmaceutical research will be described from drug discovery through to drug approval, and the strict sequential nature of the investigations will be highlighted. Next, the evolution of the linear development process will be described from a regulatory perspective. The current regulatory structures and processes will be discussed while the structure and hierarchy of the current guidance documents that increasingly drive the design and analysis of clinical trials will be presented from a statistics perspective. Furthermore an insight into the current regulatory climate will be provided with regard to the representation and evaluation of specific subject types. Finally in this chapter, the randomised and controlled clinical trial (RCT) - which underpins the confirmatory phase of drug development - will be introduced. The roles of sub-populations and subgroups are intrinsically interwoven with the properties of the RCT, and a description of the basic principles of this design is key to explaining the direction of this research effort and the importance of data sub-setting in the evaluation of robustness.

1.2 DRUG RESEARCH

1.2.1 Drug discovery

At the earliest stage, pharmaceutical research begins with drug discovery. That is, the identification of lead compounds that show specific biological activity indicative of therapeutic potential. One approach to drug discovery is mass screening - a somewhat indiscriminate process whereby a large number of compounds from the general inventory are subjected to routine tests of biological activity. An alternative more direct approach is targeted screening. In this case, compounds are selected or prepared for specific screening based on an understanding of disease intervention leading to judgements about which classes of compound are most likely to have biological activity (Schultz *et al*, 1988). Once a compound with activity has been identified, attempts are made to modify the known molecular structure with the express aim of producing a more active or selective molecule. This process is called synthesis. Alternative processes include isolation from natural plant or animal tissues or the modification of microbiological fermentation (Bohidar and Peace, 1988). Later the solubility and stability of the modified compound are evaluated with the aim of establishing a form suitable for humans and consideration is also given to the scale-up process since sufficient quantities of the compound are required for use in clinical trials. The formulation process can markedly influence the performance of the compound and consideration needs to be given to both the compound itself and the inactive excipients used - such as binding agents, which are used to hold tablets together.

1.2.2 Pre-clinical drug development

To manage the risks associated with giving a new product to humans, the product must first undergo rigorous pre-clinical (or non clinical) safety testing in a number of animal species (for instance the rat, mouse, dog and monkey). In general, limited exposure in humans is permitted once short–term animal studies have been completed satisfactorily

while longer term animal studies are required to lengthen the duration of human exposure. This process is well established and structured, and study designs are usually randomised and controlled. Considerable effort is also directed towards pharmacokinetic investigations to understand the absorption, distribution and elimination of a drug in various species. Single dose studies are followed by multiple dose studies of limited duration with exposure up to 90 days. Long term studies – particularly in rodents – are typically conducted for the life span of the animal concerned, with the aim of evaluating tumour development and associated malignancy (carcinogenicity studies). Reproductive studies are also conducted to evaluate the effects on the embryo, foetus and new-born. All these *in vivo* studies are complemented by a series of short-term *in vitro* or test tube tests (Selwyn, 1988).

1.2.3 Clinical drug development

In the clinical stage, the typical chain of events starts with one or more phase I studies in relatively few healthy volunteers to establish safety limits of exposure (toxicity evaluation) and to study the distribution of the drug within the body (pharmacokinetic evaluation). Characterising the absorption and subsequent elimination of the drug from the human body has important implications for the selection of dose levels and dosing frequencies in later clinical trials. Researchers typically plan dosing schedules with the express aim of avoiding unwanted peaks in drug concentrations since toxic effects will often be related to the presence of excessive amounts of drug in the body. However a balance needs to be maintained between the management of toxicity and the desire to achieve the required therapeutic effect (with potentially high drug concentrations). Treatments for which only a restricted range of drug concentrations achieves the required balance of therapeutic benefit with acceptable toxicity are said to have a narrow therapeutic window. Indeed in some cases a window is never established and the development of the new drug is stopped.

As a general rule, the concentration of the drug at the site of action determines the strength of effect, and drugs act by either stimulating or blocking receptors on or within cells. (In some cases it is the cells of bacteria or parasites that are targeted - for instance, drugs used to treat infectious diseases act in this way.) Drug effects are usually reversible and diminish as the drug is eliminated from the body - most commonly via the liver (hepatic) or kidneys (renal). At a particular point in time, the balancing effects of actual drug absorption and elimination determine the drug concentration within a subject, and variability exists both within and between subjects. Within a subject, external factors such as food, pregnancy, physical activity and concurrent drugs can have a significant impact on drug concentrations, while genetic factors most notably affect drug concentrations between subjects through differences in drug metabolism. These differences between subjects in the metabolism of drugs are typically due to differences in enzyme expression - indeed in some cases subjects may not express a particular enzyme at all. The impact of genetic variation is most obvious when differential effects are associated with physical characteristics such as race and gender. Other common factors that can impact drug concentrations are age, weight and concurrent disease. Indeed the doses of many treatments are calculated on the basis of body weight to account for this specific source of variation - a feature that is particularly relevant to the treatment of paediatric subjects.

After completing phase I, the next step is to administer the drug to a modest number of patients (that is, subjects with the disease) and to evaluate efficacy alongside safety for a limited period of exposure. These phase II studies will usually evaluate more than one dose level of the test drug and will include one or more control groups to establish relative dose effects. However efficacy endpoints in these studies may simply be surrogates for long-term target outcomes - such as increased survival - and follow up is often short-term.

Once signs of efficacy are established, the clinical trial programme moves to the confirmatory phase (phase III) where the challenge is to produce - through the generation of a substantial amount of high quality data - robust evidence to support the approval of the drug in specific disease indications. It is to these large studies that most clinical regulatory guidance is directed and as such the design, conduct and reporting of these studies must satisfy strict requirements. These studies typically have a broad geographic spread of investigator sites and less restrictive entry criteria compared with phase II trials - not least to be able to recruit the substantial number of subjects required. These studies also present the first real opportunity to investigate differential treatments effects across subgroups of subjects, and in some cases these investigations will be driven by previously observed differences in the pharmacokinetic profile of the drug.

1.3 DRUG REGULATION

1.3.1 History of drug regulation

Beginning with the Pure Food and Drug Act in 1906, drug regulation has been driven primarily from the US, and although many countries and geographic regions have developed their own regulations, most are based upon the guiding principles established in the US over the last 100 years or so. The 1906 Act was actually a response to growing concerns in America about the practice of adulteration and misbranding of food and drugs - a concern which centuries earlier had led to the first food law in England. (*The Assize of Bread* was proclaimed by King John in 1202 and prohibited the adulteration of bread.) However rather than being directed at drug approval, the primary focus of the Act was drug labelling with the resulting prohibition of interstate commerce with regard to unlawful food and drugs. In this respect, drugs that did not conform to documented standards of strength, quality and purity could only be sold if the specific variations were

clearly stated on the label. Furthermore the label could not contain false or misleading information. The actual requirement for regulatory approval prior to marketing came into force much later with the 1938 Food, Drug and Cosmetic Act. This act obligated companies to prove the safety of new drugs before they could be sold, although it was an amendment to the Act in 1962 (Kefauver-Harris Amendments) which eventually added the requirement for proof of efficacy (FDA History, 2002). Importantly this amendment also included the requirement to submit *substantial evidence* to support regulatory approval and for this to originate from *adequate and well-controlled investigations*. In many ways this was the defining moment for statistics in drug development, and as Anello (1999) states, *More than any other law or regulation, this law made sound statistical methodology an integral part of the regulatory process*. In the modern era of drug regulation, the thorough statistical evaluation of clinical trial data plays a crucial role in determining both drug approval and the subsequent label content since all claims regarding a drug must be substantiated. Importantly pharmaceutical companies may only market drugs within the limits of the agreed label – although individual physicians are still at liberty to prescribe an approved drug outside of these limits. (This so-called off-label usage is a particular concern for the treatment of children since often no evidence regarding the safety and efficacy of drugs approved for the treatment of adults is available for the paediatric population.)

Overall, the history of drug regulation can be viewed as a series of responses made to either notable tragic events - such as thalidomide related birth defects (1962 Kefauver-Harris Amendments) - or to growing concerns regarding the activities of the pharmaceutical companies - such as a reluctance to conduct paediatric studies (2002 Best Pharmaceuticals for Children Act). These responses have now become the embodiment of the regulatory process. However, it is also important to note that regulatory requirements

are not limited to establishing that a new drug is safe and effective for intended use prior to regulatory approval. Considerable control is also directed toward pre-approval activities with the aim of protecting the subjects who participate in clinical research and who are the source of the evidence contained within the regulatory submission for approval. For instance, drugs must be thoroughly tested in animals prior to being tested in humans, subjects must give their informed consent prior to entering a clinical trial, and regulatory authorities must approve the use of new drugs in clinical trials. An integral part of the commitment to protect subjects enrolled in clinical trials is the regular reporting of data to the regulatory authorities from ongoing clinical trial programmes.

From a US perspective, Chow and Pong (1998) give an overview of the regulatory approval process while Johnson (1988) provides a more detailed history of drug regulation. Pocock (1983) describes the corresponding developments in the UK.

1.3.2 Current regulatory structures and guidance

The most influential regulatory body is the world is the US Food and Drug Administration (FDA). From humble beginnings in 1862 when a single chemist – Charles Wetherill - was employed by the Department of Agriculture it has grown to have over 9000 multidisciplinary staff in 2002. Regulatory functions were added to the FDA in 1906 with the passing of the Food and Drugs Act, and these complemented the original scientific undertakings leading to what is now regarded as the modern era of the FDA – although the current name was not established until 1930 (FDA History, 2002).

For a new drug to be used in a clinical trial, a pharmaceutical company must first submit an IND (Investigational New Drug) application to the FDA and unless otherwise notified may commence a trial within 30 days of receipt of their application. A central feature of

the IND application is an account of the proposed clinical trial program but detailed information regarding the evidence acquired to date to support the experimentation in humans must also be provided. The requirements apply for the complete duration of the clinical trial program and effectively permit the collection of efficacy and safety data to support a NDA (New Drug Application) whilst providing exemption to the company from the law that disallows interstate commerce for an unapproved drug (Chow and Pong, 1998). Once the agreed clinical trial program is complete and if the data are believed to support the approval of the new drug, then the pharmaceutical company submits the NDA to the FDA for review.

In Europe, a new centralised system of drug regulation was implemented in 1995 with the creation of the European Agency for the Evaluation of Medicinal Products (EMEA) based in London, UK. The EMEA is essentially the focal point for a network agency based on co-operation amongst the national competent authorities of the Member States of the European Union - for example, the UK's Medicines and Healthcare products Regulatory Agency (MHRA) - formerly known as the Medicines Control Agency (MCA). In this respect, the EMEA co-ordinates the scientific resources made available. The principal scientific body of the EMEA in relation to human medicines is the Committee for Medicinal Products for Human Use (CHMP) and the EMEA is able to support the European Commission regarding harmonisation tasks within Europe and between regions at the international level. (Note that the CHMP was previously named the Committee for Proprietary Medicinal Products (CPMP) and most of their guidelines continue to use the CPMP acronym. As a result the CHMP will mostly be referred to as the CPMP within this research thesis to avoid confusion.)

The European Union Directive 2001/20/EC implemented on 1st May 2004 simplified and harmonised the process across Member States in Europe with regard to the administrative procedures governing clinical trials and in particular the implementation of Good Clinical Practice. In terms of clinical trial initiation, pharmaceutical companies (The Sponsor) are now required to submit a clinical trial authorisation (CTA) prior to commencement and each trial is assigned a EudraCT number so that it may be entered on the a clinical trial database that contains information on all interventional clinical studies of medicinal products in the EU (MHRA, 2004). An initial assessment is undertaken within 30 days but all trials must also receive ethics committee approval prior to commencement. The harmonisation effort has been far reaching and, for instance, prior to implementation of the Directive, healthy volunteer studies in the UK were actually unregulated. If a CTA is awarded the Sponsor is responsible for reporting serious unexpected adverse events (SUSARs) to the regulatory authorities and relevant ethics committees (one per country in which the trial is undertaken) within an agreed timeframe and provide annual safety reports.

In terms of drug evaluation, the European system provides two potential routes. Firstly, the centralised route, whereby applications are submitted directly to the EMEA. A scientific committee then performs the evaluation that is then transformed by the Commission into a single authorisation applying the European Union as a whole. Secondly, the decentralised procedure that is based on the principle of mutual recognition of national authorities. In this case marketing authorisation granted by one Member State is extended to one or more other Member States identified by the applicant. In cases of disagreement between Member States then the EMEA acts as arbitrator. Single national authorisations still remain an option to pharmaceutical companies, however.

Most importantly however in the context of global drug development, there has been a major initiative over recent years to harmonise regulatory requirements across regions - in particular involving the US, Europe and Japan. The so-called International Conference on Harmonisation (ICH) process has brought together representatives from regulatory authorities and experts from within pharmaceutical industry with the express aim of making recommendations regarding the technical requirements for new pharmaceutical products. The first conference was held in Brussels in 1991 (ICH 1) while the sixth meeting took place in Osaka, Japan in November 2003 (ICH 6). In the way that regulation in the US evolved from the requirement to ensure consistency from State to State, ICH is an attempt to ensure consistency between regions. The recently introduced Common Technical Document (CTD) is a prime example of this harmonisation effort (ICH M4, 2000). Pharmaceutical companies use the CTD to summarise their evidence on a new treatment to regulatory authorities and it represents the culmination of research effort, costing in the region of $800m. The CTD provides companies with a common structure and format to submit their evidence and replaces the previously tailored approaches required by specific countries or regions. The CTD is modular in format and its breadth is extensive - although the actual content requirements can still differ somewhat between countries and regions. The CTD became mandatory on 1st July 2003.

The chief benefit of the ICH process has been the generation of guidelines covering a wide variety of topics relevant to drug development. Regional authorities such as the FDA, CHMP and Japan's Ministry of Heath and Welfare have then adopted these ICH guidelines leading to a common set of standards. (In the case of the FDA, these guidelines are incorporated into the Federal Register.) The most pertinent document from a statistics perspective is the ICH E9 guideline entitled *Statistical principles for clinical trials* which was adopted by the CPMP in Europe in March 1998. (Other ICH documents that are

directly relevant to statisticians are: E3, *Note for guidance on structure and content of clinical study reports* (CPMP/ICH/137/95); E6, *Guideline for good clinical practice* (CPMP/ICH/135/95); and E10, *Choice of control group in clinical trials* (CPMP/ICH/364/96).) ICH E9 specifically targets the design, analysis and reporting of clinical trials, although many other ICH guidelines also refer to statistical issues and contain related text in assorted sections. (The content of ICH E9 was heavily influenced by the earlier *Note for Guidance* (III/3630/92-EN) produced by the CPMP in 1994 entitled *Biostatistical Methodology in Clinical Trials in Applications for Marketing Authorisations for Medicinal Purposes* (Lewis *et al*, 1995) Other relevant contributing documents included the FDA's 1988 *Guidance for the Format and Content of the Clinical and Statistical Sections of a New Drug Application* and the Japanese Ministry of Heath and Welfare's 1992 *Guideline on the Statistical analysis of Clinical Studies*.) In Europe, ICH E9 is supported by a series of CPMP points to consider documents (termed guideline in the case of the most recent issue) that address specific statistics topics in more detail while many of the therapeutic points to consider documents also address statistical issues specifically related to their area. The current list of points to consider documents that address statistics topics includes the following documents:

Reference	**CPMP Document**	**Adopted**
CPMP/EWP/908/99	Points to Consider on Multiplicity issues in Clinical Trials	Sept 2002
CPMP/EWP/2863/99	Points to Consider on Adjustment for Baseline Covariates	May 2003
CPMP/EWP/1776/99	Points to Consider on Missing Data	Nov 2001
CPMP/2330/99	Points to Consider on Application with 1.) Meta-analyses and 2.) One Pivotal study	May 2001
CPMP/EWP/482/99	Points to Consider on Switching between Superiority and Non-inferiority	Jul 2000
CPMP/EWP/2158/99	Guideline on the Choice of Non-Inferiority Margin	Jul 2005

There is one other relevant document in preparation.

Reference	CPMP Document	Status
CPMP/EWP/2459/02	Concept Paper on the Development of a CPMP Points to Consider on Methodological issues in Confirmatory Clinical Trials with Flexible Design and Analysis Plan	Concept
CHMP/EWP/83561/2005	Guideline on Clinical Trials in Small Populations	Draft

The ICH E9 guideline represents an important development in the design and analysis of clinical trials since for the first time, world-wide regulatory expectation with regard to statistically related issues has been clarified. This has the potential to raise standards and promote consistency whilst providing pharmaceutical companies with a range of issues that they should address proactively when planning the design, analysis and reporting of clinical trials. Within the regulatory framework there are still, of course, opportunities to discuss planned clinical trial designs with regulatory staff - including statisticians - although this process is more formalised in the US than elsewhere - particularly with regard to statistical advice (Lewis, 1995). In the US, the statistical review of new drug applications has been at the heart of the drug approval process and a substantial number of statisticians have been directly employed at the FDA for many years. Indeed statisticians at the FDA will actually perform their own independent analyses on the data presented in the CTD. However European authorities have historically been much slower in recognising the need for permanent statistical expertise and in fact, in the UK, the MHRA did not recruit its first statistician to the Licensing Division until 1994 (Lewis, 1996) after receiving pressure from a group of eminent medical statisticians in the UK (Pocock *et al*, 1991). (The Committee of Safety of Medicines, that advises the UK licensing authority, and the Medicines Commission, that advises government ministers in relation to the

Medicines Act of 1968, have traditionally had eminent academic statisticians as members however (Lewis, 1996)). Sweden and Germany employed permanent statistical staff earlier than this but today many countries still rely on external experts to provide advice when required. Independent statistical analysis by regulatory authorities in Europe is therefore restricted to simple analyses of transcribed data and it is more likely that a European regulatory authority will ask the pharmaceutical company to undertake additional analyses on their behalf.

Notwithstanding the availability of detailed guidance documents to aid pharmaceutical companies in the design and analysis of clinical trials, a number of papers have been published recently describing statistical shortcomings and issues in licence applications. These include articles originating from statisticians employed by the MHRA (Lewis, 1995; Lewis, 1996; Lewis and Facey, 1998; Brown, 2003), FDA (Anello, 1999), Germany's Federal Institute for Drugs and Medical Devices (Röhmel, 1999), and all regions combined (Lewis *et al*, 2001). Of note is that the recurrent themes and areas of concern include the topics covered in this research thesis - that is, the choice of analysis populations, subgroup analyses, multiplicity and the interpretation of equivalence trials. In contrast, Pong and Chow (1997), Phillips *et al* (2000) and Phillips and Haudiquet (2003) highlight the practical issues and challenges of applying the ICH E9 guideline from the pharmaceutical industry's perspective.

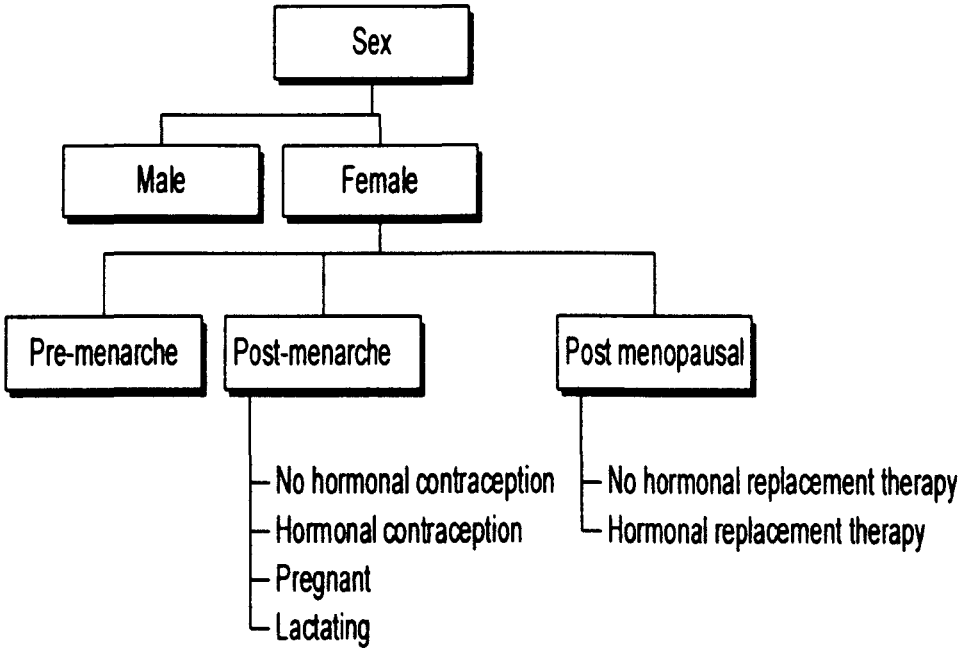
1.3.3 Current regulatory climate

In recent years the regulatory climate has changed somewhat. Following drug failures such as thalidomide, drug regulation had been directed towards limiting the exposure of subjects to the effects of experimental treatments but in recent years this *age of protectionism* has been replaced by the *age of inclusionism* (Johnson-Pratt and Bush,

1996). That is, an emphasis on the broadening the inclusion criteria for clinical trials. Interestingly the main driving force for this change was the perception that specific subject types were being denied the potential benefits from medical research, while it was also clear that the information available regarding the effects of drugs in some sections of the community – particularly children - was simply inadequate.

The focal point of the debate on adequate representation was the inclusion of women in clinical trials and the widely held view was that the treatment of women in society was actually based on evidence from clinical trials with men. Concerns were identified in relation to both the female physiology (such as the impact of the menstrual cycle and menopause [Refer to Figure 1.1], and of factors such as weight, fat content and general hormone levels) and possible interactions with other drugs (such as oral contraceptives and hormone replacement therapy).

Figure 1.1. Some lower level subgroups within the female subgroup.



Indeed it was certainly true that women of childbearing potential had systematically been excluded from early phase clinical research in all but the most severe diseases – predominately as a means of protecting the foetus and future reproductive potential (Bennett, 1993). There was also increased awareness of the more general need to individualise treatments for the optimum treatment of subjects, and to consider dose adjustment on the basis of factors such as concurrent disease, subject characteristics and concomitant medications. In this respect the emphasis began to shift towards the investigation of the resulting data from clinical trials rather than adequate representation *per se*.

In 1993, the FDA issued specific guidance in relation to women in clinical research, namely the *Guideline for the Study and Evaluation of Gender Differences in the Clinical Evaluation of Drugs*. This guideline stated the FDA's revised expectations regarding both the representation of women in clinical trials and the subsequent investigation of the data generated for potential between gender differences – including an assessment of potential differences in the pharmacokinetic profile. In addition, the FDA modified its 1977 policy that had effectively excluded women of childbearing potential from early phase trials. The rationale for the change was the view that risks to the foetus could be actively managed in early phase studies. For instance, these studies often require just a single dose of drug that could be administered after a negative pregnancy test and contraception usage would be concurrent. Furthermore it was noted that if gender differences could be demonstrated early on with regard to the pharmacokinetic profile of the drug in these studies then later phase studies could be designed more effectively. In a similar vein, the NIH Revitalisation Act was signed in the US in 1993 and included a requirement that NIH trials were *designed and carried out in a manner sufficient to provide for a valid analysis of whether the variables being studied in the trial affect women or member of minority groups, as the case*

may be, differently than other subjects in the trial (NIH Revalisation Act, 1993). The consequences of this Act led to considerable debate within the US scientific community and in particular with regard to the perceived constraint that future trials required the female quotas. However it was later shown by Meinert and Gilpin (2001) that despite the concerns that had led to the Act, over the period 1975-1995, there had actually been a sizeable excess of female-only trials and the perception that women had been understudied was misrepresented. (Merkatz *et al* (1993) give more detail of the sex specific issues in drug development and give a range of examples where differences have been observed between male and females with regard to specific drug effects.)

Similarly, earlier concerns had been expressed regarding the representation of elderly subjects in clinical trials and the requirement to investigate potential differences in drug effects between older and younger subjects. As a result the FDA issued the *Guideline for the Study of Drugs Likely to be Used in the Elderly* in 1989. In 1995, the FDA subsequently set forward a broader requirement for the presentation of effectiveness and safety data for the demographic characteristics sex, age and race and further subgroups defined by other pre-randomisation factors such as disease severity and renal impairment (Federal register, 1995). The FDA stressed that the aim was not to require the enrolment of specific subject numbers in individual studies or even the clinical trials program as a whole – rather the focus would be on the presentation of the data collected. Indeed the reference to a program of clinical trial trials is important since the FDA requires a drug application to include so-called integrated summaries of efficacy (ISE) and integrated summaries of safety (ISS). These summaries combine data from similar study designs included in the clinical trial program and are aimed at investigating consistency of effect across important subgroups. With the advent of the CTD these remain a FDA requirement although the emphasis has moved to describe these summaries move accurately as

integrated analyses. To date, formal summaries across studies are not an EMEA requirement although it is unlikely in practice that these will be removed from the CTD for European drug submissions.

Most recently regulatory focus has been directed towards paediatric clinical trials, or rather the lack of such trials. Again regulatory developments in this area have been driven from the US, and the effectiveness and safety of new treatments must now be established in the paediatric population if the product is likely to be used in a substantial number of children or if meaningful therapeutic benefit over current treatments is likely. Indeed standard labelling for drugs often prescribed for children such as "safety and effectiveness in paediatric patients have not been established" is now considered unsatisfactory following the rules that took effect in April 1999.

The ICH process has also addressed specific subject types, and has issued their own guidelines in the areas of geriatrics, paediatrics and ethnic factors although interestingly no guideline directed towards gender has ever been issued. Details of these guidelines are given below:

CPMP Reference	**ICH Document**	**CPMP Adopted**
CPMP/ICH/289/95	E5: Ethnic Factors in the Acceptability of Foreign Clinical Data	Mar 1998
CPMP/ICH/379/95	E7: Clinical Trials in Special Populations: Geriatrics	Mar 1994
CPMP/ICH/2711/99	E11: Clinical Investigation of Medicinal Products in the Pediatric Population	Jul 2000

With regard to representation, the ICH guideline for geriatrics is more prescriptive than the FDA guidelines and actually recommends a minimum target of 100 elderly subjects

(defined as 65 years or older) to allow for the detection of clinically important differences. It also encourages the inclusion of subjects aged 75 years or older and the avoidance of upper age cut-offs. However, apart from the definition of an elderly subject and the sub-division of paediatric subjects into five subgroups, there has been little direction from the regulators regarding how the effectiveness and safety data should be summarised or analysed.

1.4 THE RANDOMISED AND CONTROLLED CLINICAL TRIAL

1.4.1 Randomisation and control

In simple terms a clinical trial is a planned experiment with the essential characteristic that *one uses results based on limited sample of patients to make inferences about how treatment should be conducted in the general population of patients who will require treatment in the future* (Pocock, 1983). However it is the introduction of the randomised control that brings true scientific rigour to the clinical trial and which creates the framework for reliable interpretation. Concurrent control is essential in the confirmatory setting since within the defined experimental environment it enables an observed effect to be attributed to the experimental treatment alone - rather than to the impact of one or more concurrent external factors. It is then the instrument of randomisation that ensures that experimental and control treatments are allocated without bias and provides the probabilistic basis for treatment comparisons. Indeed it is the method of choice to control for potential confounding factors since random assignment provides the basis for causal inference (Breslow, 2001).

RA Fisher developed the randomised experiment between the World Wars, although ultimately it is Bradford Hill who is credited with introducing the concept to clinical research in the 1950's (Silverman, 1991). The randomised clinical trial is now firmly

embedded within medical research culture where it has been described by some as the most important scientific advance of the last century – a view supported by the fact that over ¼ million randomised clinical trials had been published during this period (Harrington, 2002).

Randomisation is widely held to have three important functions in the clinical trial setting. It provides protection against selection bias in the assignment of treatment to subjects by the Investigator. Over all randomisations, it generates treatment groups that are balanced with regard to factors - known and unknown, measured and not measured - that influence outcome but are independent of the treatment assignment (Gillings and Koch, 1991). (Note that a treatment group can also refer to a sequence of treatments. Indeed cross-over designs (Senn, 1993), where treatments are compared within rather than between subjects, are of particular importance in early phase clinical development.) Finally, given that the randomisation procedure was not violated, it enables test statistics to be generated for treatment comparisons (Fisher LD *et al*, 1990). Some authors regard protection against selection bias as the most important property of randomisation (Fisher LD *et al*, 1990) while others choose its *probabilistic basis for comparability* (Koch and Sollecito, 1984). Harrington (2002) describes the RCT as *one of the delightful ironies of modern science* since the action of introducing chance variation into a strictly controlled experiment provides the very means of accounting for both observed and unobserved heterogeneity. For further discussion on reasons to randomise – in particular in relation to blinding - refer to Senn (1994b).

1.4.2 Blinding

Although randomisation affords an unbiased assignment of treatments to subjects, other important sources of bias remain and these must be carefully managed to ensure that

treatment comparisons are not distorted. Knowledge of the treatment assignment on the part of the subject or the investigating team has the potential to influence the conduct and interpretation of a trial markedly. For instance, this information can influence compliance, dose modification decisions, study continuation, the recording of adverse events, the assessment of efficacy and the use of other medications; it can even influence the recruitment of further subjects into the study (Pocock, 1983). However, susceptibility to these different forms of unintentional and intentional bias will vary from study to study. For example, compliance will be less of an issue in a single dose study compared with a 12-month study with dosing three times a day. Similarly a hard outcome such as survival in oncology will be less easily influenced than the completion of a psychological measurement scale in an area such as depression. To address these sources of potential bias it has become standard practice to mask as many persons involved in the conduct of a clinical trial, as is practically possible. This design feature is called blinding and a variety of terms have been introduced to describe the degree of blinding implemented and maintained during a study. In terms of minimising bias, double-blind masking is the optimal approach since all persons involved in the conduct of the study are unaware of the treatment assignment for the complete duration of the study. Single-blind refers to cases were the subject is unaware of the treatment assignment but the investigator is unblinded - although it not uncommon to have, in addition, blinded central evaluations of outcome based on source documentation in these circumstances. Finally open-label refers to cases where both the subject and investigating team are unblinded to treatment and again blinded central evaluations may feature. It is also standard practice to mask the data management and statistics teams until such time that the database is signed off as consistent and complete.

In many cases, blinding is a matter of practicality - although regulatory authorities expect pharmaceutical companies to be implement the highest degree of masking possible or address issues of unblinding with alternative procedural control. (For example, to minimise selection bias in an open-label study - comparing an oral drug with an intravenous comparator, say - a central randomisation procedure might be used whereby the investigator must first register the subject as eligible for inclusion in the study before receiving the treatment assignment.) The exact nature of the masking challenge depends upon the study hypothesis being tested and can be broadly divided into studies designed to investigate absolute or incremental treatment effects compared with studies designed to investigate relative treatment effects. In the former case, these studies are designed to show the superiority of a test treatment versus no treatment and effective blinding requires the use of the so-called placebo control. A placebo is identical in appearance and taste to the corresponding test drug but contains an inert substance that has no pharmacological effect. As such subjects are randomised to either test treatment or placebo control, and the subjects who are assigned to placebo follow an identical regime as the subjects assigned to the active compound. Similarly add-on studies compare a test treatment to a placebo in the presence of another treatment that both groups of subjects receive concurrently (Senn, 2002). In contrast, studies designed to investigate relative treatment effects involve an active control treatment where the hypothesis is either to test for superiority of the test treatment over the reference treatment, or alternatively to demonstrate non-inferiority (or equivalence) of the test treatment versus the reference. In these cases blinding is more of a challenge since it is difficult to make a different active treatment look and taste the same without changing the properties of the formulation (e.g. the absorption of the drug). Although encapsulation of drugs is undertaken in practice, the most common solution is to employ a double-dummy technique whereby subjects receive both the active drug (either test or reference) and the correspondingly matched placebo. Of course, the actual

pharmacological effects of a treatment can unblind an individual subject and this is a particular problem for drugs with a unique toxicity profile. However this is perhaps more of a perceived problem and often subjects and investigators who think they know the treatment assignment are later proved mistaken. Finally, since no knowledge of the treatment assignment is required to create a conclusion of equivalence - this can be achieved by simply assigning a random response to all subjects, regardless of treatment allocation - the real power of randomisation and blinding is to strengthen a conclusion that the treatments differ (Senn, 1991).

1.4.3 Stratification, multiple centres and blocking

An unrestricted randomisation that assigns subjects to test treatment or alternatively to a control in equal proportions simply allocates subjects on the basis of a coin toss. (Unequal assignments are also used although ratios in excess of 2:1 are relatively uncommon due to an accelerated increase in the type II error.) However, in practice most designs incorporate some measure to limit the chances of a large imbalance occurring in the realised randomisation and random assignment is often restricted by the use of blocks and stratification. Fixed blocks contain a strict sequential number of random assignments such that within each block the treatment assignment ratio is enforced. For example, a block size of six for a trial that plans to randomise an equal number of subjects to each of two treatments would include 3 assignments per treatment group. Blocks are therefore a method of facilitating balance on an ongoing basis by ensuring that the proportion of subjects assigned to each treatment is close to the planned proportion, no matter if the trial is stopped early or randomises more subjects than planned. In multi-centre studies it is common practice to assign multiples of complete blocks to each centre that in effect stratifies the study by centre. Stratification is a method of ensuring that the proportion of subjects assigned to each treatment within each pre-defined stratum is close to the planned

proportion, and is equivalent to creating a separate randomisation schema for each stratifying factor or stratum. Stratification can also be used to address representation. That is, it can be used to ensure that specific quotas of subjects with a pre-randomisation characteristic - such as gender or disease severity - are randomised into the study.

One important implication of restricted randomisation is that all factors used to stratify the design - including study centre, where appropriate - should be included in the statistical model in the subsequent analysis (ICH E9, 1998). Inclusion of the blocking factor itself in the model is rare and is not a regulatory requirement. In cases where there are few subjects randomised per centre, then centres are often grouped together by country (or by using some other pre-specified rule).

1.4.4 Eligibility

A clinical trial protocol defines the intended study population through detailed inclusion and exclusion criteria. These criteria are applied at a screening visit prior to randomising the subject into the study and in some designs there is a screening or run-in period where subjects must satisfy criteria at both entry and completion of this period. Screening periods can also be used to determine potential compliance; withdraw current treatments or add new ones, train subjects and investigators in terms of specific procedures or tests; provide additional baseline measurements to control variability; and determine within subject variability.

Inclusion criteria can generally be regarded as referring to factors that affect the assessment of efficacy. Typically these criteria define the presence and severity of the disease under study. They will also be used to define the broad population under study in terms of the demographics, age and sex, and as such sets out the opportunities for different

subject types to be represented in the study. Agreement to participate in the study by the subject is also included here – that is, informed consent. In contrast, exclusion criteria are generally specified to avoid the inclusion of subjects who could be at risk from any of the study treatments. Typical examples include: pregnant women or women not using a suitable form of contraception; subjects with hepatic or renal dysfunction for whom safety data are not yet available; subjects taking concomitant medication for which there are no data on the possibility of interaction with one or more of the study treatments; and concurrent illnesses which may make the assessment of safety and efficacy difficult. One important feature of eligibility criteria is that they are usually applied prior to randomisation and as such are independent of the treatment assignment. The exclusion of ineligible subjects - who have been randomised in error - from subsequent analyses on the basis of not meeting these criteria cannot introduce bias in terms of treatment group comparability. However it should be noted that unduly restrictive inclusion and exclusion criteria risk selecting a trial population for study that is no longer representative of the intended target population for treatment. In this respect the danger would be that an unbiased estimate of the wrong parameter would result.

1.5 DISCUSSION

Drug development is essentially a well-established, step-wise process whereby knowledge is accumulated through time until such time that the aggregated evidence either supports drug approval or alternatively the discontinuation of development activity. Increasingly, the world of drug regulation is becoming a more standardised and well-defined environment, where the application of statistical methodology and thinking plays a key role in design, analysis and interpretation of clinical trial programs. Accordingly specific statistical guidelines have been created to define regulatory expectation although, according to the regulators, problems remain and many submissions fail to address key

issues in a satisfactory manner. Alongside the developments in statistical guidance, the regulatory focus has been on inclusionism, and the identification of differential treatment effects across subgroups with the aim of determining the need for dose adjustment in cases where a drug is not uniformly safe and efficacious in a particular disease area. In this respect robust evidence is key to a successful regulatory submission and it is perhaps no coincidence that the regulatory statisticians have highlighted the analysis and reporting of subgroups and sub-populations as particular problem areas. Therapeutic equivalence has also been identified as a problem, and specific issues relating to analysis populations and subgroups are also relevant here. Drug development has been sorely neglected in the paediatric population and drug labelling has more frequently excluded their use in children rather than addressing the real issue that the paediatric population requires treatment options that have been shown to be safe and effective. Although the regulatory authorities are now remedying this situation, clinical trial research is not well developed and unique challenges exist which must be overcome.

In Chapter 2, the basic principles of data sub-setting will be examined in more detail. In particular, the exclusion of clinical trial data to form sub-populations and subgroups will be investigated and the impact on the estimated treatment difference evaluated.

CHAPTER TWO: (SUB)DIVIDE AND RULE(S)

Brother Tony
Was notably bony
He would never eat his meals
And so they gave him medicinal compound
Now they move him round on wheels.

2.1 INTRODUCTION

The aim of this chapter is to illustrate some of the practical challenges faced when constructing sub-populations and subgroups in the reporting of clinical trial data. Firstly, the basic principles of sub-setting clinical trial data will be discussed. The focus will be on randomisation as the basis for providing unbiased estimates of treatment effects, and it will be shown how the exclusion of data has the potential to introduce bias into the estimation process. Secondly, regulatory guidance in relation to populations and sub-populations will be reviewed with specific emphasis on the practical challenges of implementing the intent-to-treat principle. Thirdly, approaches taken to investigate treatment effects within subgroups will be discussed together with issues surrounding multiplicity and the risk of observing inconsistent results from subgroup to subgroup. Finally, some examples of biased sub-populations and subgroups will be presented from a variety of therapeutic areas.

Now, sub-populations and subgroups both involve data sub-setting, although sub-populations tend to be based on composite criteria (such as all subjects who meet all the inclusion criteria) whereas subgroups are typically constructed or partitioned based on a single criterion (such as male subjects). However despite being closely related in terms of construction, sub-populations and subgroups play different roles in the quest for robustness in support of drug approval.

The selection of subject sub-populations revolves around issues of accountability and generalisability, and is of a philosophical nature. (Accountability refers to the general desire to account or incorporate all subjects from the trial in the subsequent statistical analysis while generalisability is directed towards determining the scope of extrapolation.) The *modus operandi* is that data exclusion creates a subset of the original complete population of all randomised subjects, and the revised estimate of the treatment difference in this subset is compared, not with the corresponding complement sub-population, but with the original complete population. In this respect the approach can be considered hierarchical - purist to interventionist - whereby beginning with the all randomised subjects, ever increasing amounts of data are excluded to form sub-populations increasingly less representative of the original complete population. As a result, one can investigate how sensitive the study conclusions are to such data exclusions, and if consistent results are achieved across all analysis populations, the conclusions drawn are considered robust.

Essentially there are three main reasons why one might to exclude data from the original complete population of all randomised subjects to form a sub-population in a RCT. Firstly there may be problems associated with the randomisation procedure itself. Secondly, some subjects may have been randomised who were actually ineligible for inclusion into the study - for instance, subjects without the required disease severity. Thirdly, subjects may not - within limits - conform to the procedures of the protocol or may be prematurely withdrawn. For instance, a subject may not take the treatment as requested and may be viewed as being non-evaluable. Sub-populations that account for these types of exclusion are often referred to as per protocol (PP) populations since the subjects included are deemed protocol compliant.

Now, while the specification of various analysis populations or sub-populations is a well-accepted feature of confirmatory studies, only a single population is usually selected to secure the primary investigation of the primary efficacy endpoint - a procedure that effectively controls for multiplicity. Furthermore, the make-up of this primary population will, in most instances, be as close to the original complete population of all randomised subjects as is practically possible – a practice that ensures that results are broadly representative of the subjects taking part in the study. However, rather than direct extrapolation of the results to a precisely defined subject population, generalisation is directed mostly towards simply demonstrating that the test treatment is broadly effective in practice.

In contrast, subgroup analyses lend themselves to precisely defined treatment comparisons either within or between subgroups although they are usually considered to be of secondary importance in an individual clinical trial or of an exploratory nature. The primary aim of such analyses is to investigate consistency of effect across clearly defined subgroups either within a study or across a series of related studies. Again the approach can be considered somewhat hierarchical - an initial overall treatment comparison followed by a treatment comparison within each subgroup. If the within subgroup differences are similar, and consistent with the overall treatment difference, then the overall difference is considered robust and broadly applicable to all subgroups. Multiplicity, for instance, can be addressed through closed testing procedures that exploit the hierarchical nature of the investigations or through tests for treatment by subgroup interaction.

In the next section of this chapter, it will be shown how randomisation is fundamental to the unbiased estimation of treatment differences. Furthermore, the impact of different

types of data exclusion will be considered regarding the potential to introduce bias into the comparison of randomised treatments within analysis populations.

2.2 POPULATIONS AND SUB-POPULATIONS

2.2.1 The role of randomisation

As described in Chapter One, in terms of experimental design, over all randomisations, randomisation balances treatment groups for treatment independent factors - known and unknown, measured and not measured – that may influence outcome (Gillings and Koch, 1991). (The qualifier is included because imbalance in the distributions of factors that have no influence on outcome can be considered irrelevant to the treatment comparisons.) It follows that if treatment groups are balanced in this way, then any observed difference between the treatments groups with regard to outcome must be attributable to either the treatments themselves or random variation. Accordingly, it is in this context that a statistical analysis that includes all randomised subjects is described as unbiased. More precisely, randomisation leads to an unbiased distribution of subjects to treatment groups. However, note that the observed randomisation for a particular study will exhibit some imbalance with respect to these factors, although clearly the extent of the imbalance can only be assessed for those factors for which data have been recorded or measured. The key element is that observed imbalance is not of a systematic nature, and statistical techniques account for imbalance either unconditionally or conditionally (Senn, 1994) - the latter approach involving additional assumptions regarding model structure.

Conversely, the exclusion of randomised subjects to form a sub-population or subgroup - if not based on a mechanism independent of randomised treatment - has the potential to introduce bias through the systematic imbalance of factors across treatment groups. Accordingly, to maintain an unbiased distribution of subjects to treatment groups, the exact

nature of exclusions must be determined when forming sub-populations or subgroups. Factors generally regarded as being independent of the treatment assignment are those belonging to subjects at the time of randomisation. That is, factors for which values were recorded - or could have been recorded - up until the point of randomisation. Exclusion of subjects on the basis of these factors does not introduce systematic imbalance and the sub-population or subgroup remains unbiased in this regard.

In the following sub-sections, three well-defined areas of data exclusion are examined in relation to the impact on treatment group comparability. These areas are problems associated with the randomisation procedure; ineligibility of subjects; and procedural non-compliance post-randomisation. In a fourth sub-section, the broader impact of data exclusion on treatment effect estimation will be discussed in terms of the relationship between treatment-independent factors and outcome.

2.2.2 Exclusions based on problems associated with the randomisation procedure

Problems associated with the randomisation process are not uncommon in randomised clinical trials although these are usually limited to at most a handful of subjects per study. The error of most concern is when the actual treatment received is different from the randomised treatment, although another important concern is subjects that are randomised more than once into the same study. In the first case, the intuitive clinical approach to analyse the subject according to the treatment received is at odds with the statistical requirement to maintain an unbiased assignment of subjects to treatment. Similarly, in the second case although the clinical view might be that each entry represents a new and distinct course of treatment, the statistical requirement for independent data is violated if more than one entry is included in the subsequent statistical analysis. Examples of potentially less serious problems with the randomisation procedure are subjects who are

randomised out of sequence and subjects who are randomised from an incorrect stratum – both of which are generally considered to be indicative of poor study conduct, although selection bias remains a possibility.

In general, the primary consideration when assessing randomisation violations is to determine whether the disruption to the randomisation procedure was itself of a random nature. The introduction of another random process, by definition independent of treatment, will have no impact on the distribution of factors, and it follows that - over all randomisations - treatment groups will remain balanced. In this regard, the level of blinding employed in the study is important, since the deliberate and systematic assignment of specific subjects to known treatments by Investigators requires accurate prediction of the randomisation schema. Therefore to predict future treatment assignments in a double-blind study requires accurate prediction of both the block size and the treatment assigned to previous subjects. Regarding the block size, it is considered poor practice to specify the block length in the study protocol, although given the current tendency to adopt fixed blocks of short length, Investigators with knowledge of trial design may well be able to make educated guesses. As for previous treatment assignments, these can be revealed through the observation of unique adverse events or marked efficacy, although stochastic elements makes this less useful to the deliberate fraudster than may at first appear. Another direct route to unblinding is the revelation of pharmacokinetic data for individual subjects, and procedural steps must be taken to avoid this happening during a clinical trial. However, despite these potential chinks, most randomisation errors in double-blind studies will be independent of randomised treatment and will be a simple reflection of poor administration and study conduct. However, in open label or partially blinded studies, it is much more difficult to demonstrate that errors are independent of randomised treatment and convincing arguments are required. Indeed it is much more

important to check in these studies for missed or unused treatment assignments and for subjects randomised out of sequence, since once the treatment for the next assignment is revealed, an obvious route exists for the deliberate and biased assignment of subjects to treatment. In fact, this is where central randomisations have such an important role to play, since Investigators are required first to provide identifying information for subjects - such as gender and date of birth – prior to the randomised treatment being assigned. Furthermore, avoiding stratification by centre when using central randomisation, limits the potential for assignment prediction. In both blinded and open label studies, missed numbers may indicate subjects who were randomised but not treated, for whom no data have been provided. In this respect, it is important that all randomisation numbers are accounted for on study completion and any missed numbers queried with the investigator site.

The assignment of the incorrect randomised treatment for all or just part of the intended treatment duration is a more likely occurrence in open label studies where the drug supply may not be pre-packaged with the subject number. In general, the greater the scope for an administration error, the more likely it is that an error will occur. Indeed since open label or partially blinded studies are usually only performed where blinding is impractical - intravenous applications with different treatment schedules, for instance - additional complex administration involving a number of different parties may well be required in comparison with double-blind studies, making these administration errors more likely.

Subjects randomised more than once into the same study represent a unique challenge, although it is now common for protocols to state explicitly that this should not occur. From a statistical perspective, the issue relates to analysing data that are not independent and it is self evident that subjects who have multiple entries in the same clinical trial

represent a subgroup of subjects who have demonstrated the ability to tolerate at least one of the randomised treatments previously. The specific issue of multiple entries will be considered in more detail through the presentation of a specific example in Chapter Seven.

2.2.3 Exclusions based on eligibility

A clinical trial protocol typically defines the intended subject population for study through detailed inclusion and exclusion criteria and, prior to randomisation, an investigator is required to document that the subject meets all inclusion criteria specified whilst meeting none of the exclusion criteria.

Inclusion criteria can generally be regarded as referring to issues that affect the assessment of efficacy or outcome and typically these criteria are used to define carefully the presence and severity of the disease under study. These criteria will also be used to define the broad population under study – for instance, restrictions relating to demographic factors such as age, sex and race. Confirmation of agreement to participate in the study on the part of subject is also detailed.

In contrast, exclusion criteria are generally specified to avoid the inclusion of subjects who could be at risk from one or more of the study treatments or from the study procedures. Typical examples include: pregnant women or women not using a suitable form of contraception; subjects with hepatic or renal dysfunction for whom safety data are not yet available; subjects taking concomitant medication for which the possibility of drug-study treatment interaction cannot be excluded; and concurrent illnesses which may make the assessment of both safety and efficacy difficult.

As the inclusion criteria are used to specify the type of subject who should be studied in terms of efficacy, there is a greater rationale to exclude subjects who do not meet these criteria to form sub-populations rather than those meeting the exclusion criteria. Since both inclusion and exclusion criteria are applied to subjects prior to randomisation, they are, by definition, independent of the treatment assignment. Accordingly, the exclusion of ineligible randomised subjects (based on inclusion and/or exclusion criteria) to form sub-populations does not introduce bias in terms of treatment group comparability - although the estimate of the treatment difference is now applicable to a more narrowly defined population.

During the course of the study, it is clear that if a subject has been enrolled in error and as a result is put at risk (typically due to a subject meeting one or more exclusion criteria), then randomised treatment should be withdrawn immediately. Although it is conceivable that the opportunistic detection of this error is in some way connected to the treatment received, in practice this is unlikely. Ineligible subjects not at risk are usually allowed to continue to study completion however when errors are detected. In terms of forming an eligible sub-population (based on inclusion and/or exclusion criteria), all subjects not meeting the sub-population criteria, regardless of their study completion status, would be excluded. However the interesting challenge arises when the analysis population includes all randomised subjects since subjects withdrawn from the study following opportunistic detection will mostly likely have missing data on the key endpoints. This point regarding missing data is discussed further in Section 2.3.3.

2.2.4 Exclusions based on evaluability

In addition to specifying the inclusion and exclusion criteria for entry into a clinical trial, the study protocol also details the procedures and expected study conduct for subjects once

they have been successfully randomised into the study. In this respect, protocol violation refers to procedural non-compliance following randomisation, and may warrant data exclusion. Accordingly, a subject who does not follow the protocol as planned is deemed non-evaluable. Typical examples of procedural non-compliance include: taking too much or too little of the randomised treatment; taking proscribed concomitant medication; missing scheduled visits; and premature study withdrawal. Since these violations occur after randomisation, their relationship to randomised treatment is uncertain, and data exclusion based on these events is therefore liable to create systematic imbalance of treatment groups and introduce bias. Indeed it is exclusions of this nature that raise the most concern for regulatory authorities.

Perhaps the most fundamental case of non-compliance post randomisation is where a subject fails to take even a single dose of randomised treatment. Clearly the risk of this happening increases if the treatment assignment is not blinded, but it also increases the greater the duration between randomisation and planned treatment. In some instances the subject may simply die before treatment can be commenced - something that is not uncommon in serious indications such as, head injury, myocardial infarction and late stage cancer. In other instances, such as migraine, a randomised subject may not have an attack with the defined period of the study and may simply complete the study never having required treatment. Of particular importance is differential risk where randomised treatments take different times to set up or administer. For instance, photodynamic therapy (PDT) requires the injection of a photosensitive drug followed by laser treatment once the drug has reached the target tissue. In clinical trials, PDT and has been compared to more straightforward laser treatment which can be commenced without the delays associated with PDT. These delays can lead to withdrawals due to toxicity or even death before the treatment schedule is complete. Another example of delayed treatment is randomisation

on the basis of clinical signs and symptoms of intra-abdominal infection in the comparison of anti-bacterial treatments. Subsequent investigation by the surgeon may find that infection is not the cause of the condition and anti-bacterial treatment is not required. In all studies the general rule should be to randomise as close to treatment as possible but as these examples show the problem cannot be entirely avoided. Subjects withdrawing consent is a particular problem in open-label studies since although it is difficult to avoid, it has great potential to introduce bias.

An important aspect of procedural compliance is that successful study conduct is by necessity reduced to a simple dichotomy of compliance versus non-compliance for data that are frequently recorded on the continuum. For instance a subject who records taking between 80% and 120% of the expected number of doses of randomised treatment during the course of the study may be defined as evaluable, and any subject outside of this range regarded as a violator. In this respect it is easy to see that evaluability criteria can be somewhat subjective and it is for this very reason that ICH E9 states that all decisions regarding the eligibility and evaluability of individual subjects are finalised and documented prior to the unblinding treatment assignment to avoid bias.

Subject withdrawal is a special case of procedural non-compliance, and may be indicative of either the efficacy of the treatment received or the lack of it, it may reflect unacceptable toxicity while in some cases it may be completely unrelated to treatment. In all clinical trials, subjects are free to withdraw from the study at any time without reason (that is, to withdraw consent) while protocols should also state specific conditions for Investigator determined withdrawal. In most cases early withdrawal will be regarded as treatment failure and specific analyses may be planned to allow for this in the construction of the outcome. However in many cases withdrawal simply represents a case of missing outcome

data and the challenge is how to include such subjects in the statistical comparison of the treatments. Section 2.3.3 investigates this issue further. Of primary concern to regulatory authorities is the case where differential withdrawal exists in a study with regard to randomised treatment since for the reasons outlined earlier, this has the potential to introduce bias if withdrawn subjects are simply excluded form the analysis population. (Note however, that the complete exclusion of withdrawn subjects from the analysis is often seen as a regulatory requirement for submissions to Japan (Frith, 2003; Christie, 2003)).

2.2.5 The impact of exclusions beyond treatment group comparability

Although it has been demonstrated that some specific types of data exclusion have no impact on the comparability of treatment groups with respect to the distributions of baseline factors, this is not the only aspect to consider when estimating treatment effects. Another important element is the relationship between data exclusion and outcome, and in fact this is the very aspect which is investigated with subgroup analyses. To take a trivial example, if all subjects with the disease under study were excluded to form a sub-population, then although treatment groups would be balanced in the remaining subjects - over all randomisations - the expected treatment difference with respect to outcome would be zero regardless of the true treatment difference in patients with the disease.

Indeed the design of the study directly influences the estimate of the treatment difference through the inclusion and exclusion criteria that define the baseline factors for the subjects that are randomised into the study. Furthermore run-in periods have the potential to select subjects that favour one treatment over another. That is, there is a difference between a study that compares test and reference treatments under the conditions of the test treatment, and one in which test and reference are compared under the conditions of the reference

treatment. For instance, subjects may enter a run-in period during which they receive non-randomised test treatment and subjects who successfully complete this run-in will have demonstrated that they can tolerate the test treatment and consequently represent a selected cohort. (An example of this is given in Section 2.5.2.) Subjects will also be required to demonstrate procedural compliance during the run-in and this could take the form of showing the ability and willingness to take treatment according to a particularly complicated or difficult treatment regimen. Again this could favour compliance in those subjects subsequently randomised to a similar regimen. In general run-ins (including those in which subjects receive placebo) may be regarded as impacting on both the expected value and variability of baseline factors.

A specific issue arises with treatment-independent exclusions based on data recorded pre-randomisation but where the results are not known until post-randomisation. The treatment of infections is a case in point, expanded upon through an example in Chapter Seven. Essentially subjects are randomised and treated on the basis of clinical signs and symptoms of infection while a sample (urine, sputum, blood etc) is usually taken from the infected area before or at the time of randomisation – this is subsequently sent to a laboratory for investigation. However since it takes a while to grow sufficient bacteria to allow identification of the potential pathogen, the result is often not obtained by the Investigator for 48 hours or more. Even then false negative and false positive results are relatively common. Hence although the subsequent exclusion of subjects based on a negative culture result (that is, no bacteriological proof of infection) will not introduce bias in terms of treatment group comparability, the resulting estimate of the treatment difference may have limited applicability and will not be broadly generalisable to clinical practice.

In the next section of this chapter, the regulatory expectations regarding the choice of analysis populations will be discussed, and in particular the so-called intent-to-treat principle will be introduced which lies at the heart of clinical trial reporting.

2.3 REGULATORY CONSIDERATIONS AND THE INTENT-TO-TREAT PRINCIPLE

2.3.1 The evolution of the intent-to-treat principle

The application of sound statistical principles is the focal point of current regulatory guidance (ICH E9, 1998) and great emphasis is placed on the management of potential bias. In relation to analysis populations, one principle in particular is prominent, even being described as approaching the position of *sacred cow* (Armitage, 1998) – this is the so-called intent-to-treat (ITT) principle. Simply stated, the ITT principle is directed towards undertaking statistical analyses that include all randomised subjects according to the treatment groups to which they were originally randomised. That is, all subjects randomised, as randomised. According to Newell (1992), the phrase first appeared in Bradford Hill's 1961 edition of *Principles of Medical Statistics* - although Lewis (1995) claims that earlier editions of the book published in the 1950's included the term.

The ITT principle can be described as essentially a *modus operandi* for statistical analysis. It is both a means of providing an unbiased estimate of the treatment difference (in terms of the randomisation) and of ensuring that such estimates are representative of the complete sample of subjects who were randomised - and by implication generalisable to the broader population from which the sample was taken. Furthermore, ICH E9 (1998) states that *under many circumstances it may also provide estimates of treatment effects which are more likely to mirror those observed in subsequent practice.*

The ITT principle uses the arguments put forward earlier regarding randomisation to achieve the status of unbiased analysis and can be considered to preserve the benefits of randomisation. However Lewis (1995) has questioned whether ITT is truly a principle preferring the view that ITT has arisen from a set of more basic principles laid out by Bradford Hill relating to sound statistical practice. These are the principles of accounting for all subjects who were randomised in the subsequent statistical analysis, and of giving careful consideration to exclusions, withdrawals and missing data. Lewis argues that it is this set of basic principles that should be followed rather than the ITT principle *per se* that has subsumed them. Indeed Lewis and Machin (1993) describe how - rather than ITT - it was the application of these more basic principles that came to the fore in the 1970's through the work of Peto *et al* (1976, 1977). These basic principles were directed to the disease area of oncology, and in particular to mortality studies where the long-term follow-up of subjects was potentially problematic. Later, it became standard practice to include all randomised subjects in the analysis of these studies, when comparing treatments with regard to mortality, and to include all recorded deaths - even those occurring after treatment completion in the follow-up phase of the study. However the application of these basic principles was not restricted to oncology and under the umbrella term intent-to-treat, they also became widely accepted throughout the 1980's and early 1990's in other disease areas - including anti-infectives (British Society of Antimicrobial Chemotherapy, 1989) and hypnotic drugs (CPMP, 1992). Unfortunately, the simplicity and compelling properties of the ITT principle did not necessarily transfer well to other disease areas, and in particular the relative ease in which survival status could be obtained and analysed in oncology studies was not matched in other areas where complete quantitative recordings at fixed time points were required. Accordingly the statistical literature contains numerous papers - including Fisher *et al* (1990), Feinstein (1991),

Gillings and Koch (1991) and Ellenberg (1996) - that discuss the practical challenges associated with the application of the ITT principle.

In 1998, the ICH E9 guideline introduced a new regulatory term in relation to analysis populations - the Full Set. The Full Set (FS) concept was a return to the more basic principles of Bradford Hill and its introduction most likely reflected the influence of Lewis in the development of ICH E9 – by now a Statistician at the MCA (now called the MHRA). The FS is defined *as the set of subjects that is a close as possible to the ideal implied by the ITT principle. It is derived from the set of all randomised subjects by minimal and justified elimination of subjects*. Notably, Lewis was also highly influential in the development of the European (CPMP) forerunner to ICH E9, where these basic principles were again to the fore. For instance, this Note for Guidance (III/3630/92-EN, 1995) stated that *decisions concerning the analysis populations should be guided by the principles underlying the 'intention to treat' and 'per protocol' strategies*. Indeed it is interesting to note the difference in emphasis for ITT in this document - that is, strategy (i.e. plan) rather than principle (i.e. dogma).

However, despite broad acceptance of the qualities and associated benefits of the ITT principle, the evolving research areas of therapeutic equivalence and non-inferiority in the 1990's led to regulatory concerns being raised with its application in both of these specific areas. These concerns were documented in both the CPMP Note for Guidance (1995) and ICH E9 (1998). In contrast to superiority trials, where the aim is to show that test treatment is superior to reference treatment (usually placebo), equivalence and non-inferiority studies aim to show that - within certain pre-defined margins - a test treatment is the same as (equivalent) or no worse than (non-inferior) an active control. This switch of

objective was the source of the regulatory concerns and the choice of analysis population was regarded as having important implications for the control of potential bias.

Now, from a regulatory perspective, one additional desirable property of the ITT principle is that the estimated treatment difference will tend to diminish as problems associated with the study conduct increase. (That is, problems connected with eligibility, evaluability and the randomisation procedure.) In this respect, the ITT principle is said to provide protection against over-optimistic estimates of the treatment difference for superiority trials (ICH E9, 1998), and it in effect penalises drug companies for poor study design and conduct. (Indeed Siegel (Ellenberg, 1996) has even suggested that it is actually the regulatory focus on ITT that encourages drug companies to make every effort to conduct quality trials by reducing drop-out and non-compliance.) Thus, although bias is not minimised *per se*, the direction of the bias is controlled in such a way that the estimated treatment difference is deemed conservative. Of course, while this is desirable for superiority trials, for equivalence or non-inferiority trials this property is, in contrast, anti-conservative. Indeed to reflect this point, the CPMP Note for Guidance (1995) went so far as to state that *the 'intention to treat' strategy is insecure* [for equivalence] while ICH E9 stated that *its role should be considered very carefully*. Subsequently, the CPMP introduced the specific *Points to Consider on switching between superiority and non-inferiority* (2000) with the aim of ensuring the most appropriate interpretation of data at the time of analysis. This document addresses some of the practical difficulties faced when switching objectives within the same trial and the choice of analysis population receives special attention. It recommends that the FS is the *analysis set of choice* for superiority while, in contrast to the Note for Guidance and ICH E9 guidelines, it states that the FS and PP populations have *equal importance* for non-inferiority.

In the following sub-section, consideration is given to the assumptions underlying the view that ITT is conservative from a regulatory perspective for superiority trials. A detailed evaluation of the choice of analysis populations in equivalence and non-inferiority trials is reserved until Chapter Five

2.3.2 The conservative nature of ITT

As described earlier, it is widely accepted that problems associated with study conduct tend to diminish the estimated treatment difference. Indeed it can be said to be somewhat intuitive that if, say, similar proportions of subjects in each treatment group failed to take any study treatment, the consequence of including rather than excluding these data (with an expected treatment difference of zero) would be to reduce the overall treatment difference. However to consider this phenomenon in a more formal sense, it necessary to introduce a framework where the impact of protocol non-compliance can be investigated when estimating treatment differences in all randomised subjects.

Let π_r be the probability of positive response ($x = 1$) in a population receiving a reference treatment. (Note the reference treatment could be placebo.) Similarly, let π_t refer to the corresponding probability of response for a test treatment. In each case, π_i is estimated by $p_i = x_i/n_i$, where n_i is the number of subjects randomised to each treatment, and x_i is the corresponding number of subjects with a positive response.

Now, some subjects may not adhere to the protocol which could lead to outcomes that are not truly representative of the treatment assigned. For instance, a subject that does not take the treatment as directed (treatment non-compliance) during the study may not have a favourable outcome. In this regard, if the subject would have had a favourable outcome if they had taken the treatment correctly, then the observed response can be viewed as a

misclassification, which is a false negative. Similarly, a subject who would have failed treatment if they had not taken an additional proscribed treatment (concomitant medication) would provide an observed response that represented a false positive if they responded in a favourable way to the additional treatment. To represent these two cases, let θ_i be the probability of a false negative and let φ_i be the probability of a false positive, with regard to the outcome for treatment i ($i = r, t$). (Note, this approach is adapted from Goldberg (1975), who applied this formulation to investigate medical screening techniques.)

Now, assume that the both the false negative and false positive rates are identical in each treatment group, that is, $\theta_t = \theta_r$ and $\varphi_t = \varphi_r$. Then when misclassification is present

$$E(p_i') = \pi_i(1-\theta) + (1-\pi_i)\varphi$$

where p_i' is the proportion of subjects with a positive response in treatment group i.

It follows that the expectation of the observed treatment difference is

$$E(p_t' - p_r') = (\pi_t - \pi_r)(1-\theta-\varphi) \qquad (1)$$

that is unbiased under a standard null hypothesis of $H_0 : \pi_t = \pi_r$. However under the alternative hypothesis, and under the assumption that $\theta + \varphi \leq 1$, Bross (1954) concludes that in the presence of misclassification, $E(p_t' - p_r')$ must be smaller than the true difference. This then is the basis of the regulatory argument for binary outcomes that non-adherence to the study protocol results in misclassification and reduces the magnitude of the expected treatment difference. Indeed it is this type of rationale that led Lewis and Machin (1993) to conclude *the inevitable dilution of the treatment effect in an ITT analysis*.

However if the restrictions of identical false negative and false positive rates in each treatment group are removed, such that $\theta_t \neq \theta_r$ and $\varphi_t \neq \varphi_r$, then the corresponding expected treatment difference is

$$E(p_t' - p_r') = \pi_t(1-\theta_t-\varphi_t) - \pi_r(1-\theta_r-\varphi_r) + (\varphi_t - \varphi_r) \quad (2)$$

which under the null hypothesis, $H_0 : \pi_t = \pi_r = \pi$, simplifies to

$$E(p_t' - p_r') = \pi[(\theta_r - \theta_t) + (\varphi_r - \varphi_t)] + (\varphi_t - \varphi_r) \quad (3)$$

It is clear from this formulation that the expected treatment difference may be larger or smaller than the true treatment difference depending on the misclassification rates (θ_i, φ_i) in each treatment group, although note that the expected bias is independent of π, if $(\theta_r - \theta_t) + (\varphi_r - \varphi_t) = 0$ (Goldberg, 1975). Therefore, although statistical analyses based on the ITT principle will remain unbiased in terms of the distribution of factors that influence outcome, a separate class of bias may be introduced from a different source – that is, differential misclassification.

Of course, the misclassification argument depends to some extent on the objectives of the study - that is, the true values of π_t and π_r that one is designing a study to estimate. Schwartz and Lellouch (1967) introduced the terms, explanatory and pragmatic, to distinguish between different trial viewpoints. In broad terms explanatory studies are aimed at estimating the difference between treatments with regard to the pharmacological or biological effect in a well-controlled environment, whereas pragmatic studies are aimed at estimating the corresponding effects of treatment when used in practice. (The Schwartz and Lellouch dichotomy actually extends much further than simple estimation, and its implications - expanded upon in great detail in Schwartz, Flamant and Lellouch (1980) - are indeed far reaching.) In the context of analysis populations and misclassification, the

Schwartz and Lellouch dichotomy is instructive, since a study that is regarded as pragmatic and which estimates the effect of treatment using the ITT principle, warts and all, may not actually consider all forms of violation as actually constituting potential misclassification. That is, the aim may be to estimate the effects of treatment in practice, where such levels of observed non-conformity are common.

A key element to the Schwartz and Lellouch philosophy is the expectation that individual studies will be designed and analysed from either the explanatory or the pragmatic standpoint - not both. However, Armitage (1998) takes a different view stating *that the two attitudes are likely to co-exist, and to compete for ascendancy, in any one trial*, whilst acknowledging the influence of the ITT principle by stating that *current practice leans very heavily in the pragmatic direction*. Indeed this is the case, and it is common practice in drug development to undertake statistical analyses on both ITT and PP populations for regulatory submission and to compare the resulting estimates in terms of a sensitivity analysis. In this respect, one could argue that the ITT and PP populations are actually aimed at estimating different parameters and the resulting sensitivity analysis is really a comparison of estimates of different parameters rather than a comparison of different estimates of the same parameter.

2.3.3 The practical challenges of implementing ITT

As alluded to in previous sections, although the ITT principle is straightforward in concept, its implementation is not necessarily unproblematic in some therapeutic areas. In response to these difficulties, the umbrella term ITT has come to be interpreted very broadly and now encompasses a wide range of definitions – some pragmatic, others simply convenient. By illustration, Hollis and Campbell (1999) published a survey of all RCTs reported in the *BMJ*, *Lancet*, *JAMA* and *New England Journal of Medicine* during 1997 in which 119

(48%) reports mentioned ITT analysis. Commenting on the findings of the paper, Day (2002) reflected on the different and shifting combinations of subject exclusion based on missing data, protocol violation and early withdrawal, describing the implementation of ITT as *many and varied.*

Some authors have attempted to formalise a practical definition of ITT, with perhaps the most influential in the area of drug development being Gillings and Koch (1991). Gillings and Koch were aware that while ITT had become widely accepted, its implementation had sometimes become confused. They set out to balance the clinical and statistical perspectives, and in particular focused on subjects who had not received randomised treatment or who had no post-randomisation efficacy data. They settled on the following practical definition of ITT:

> *All patients randomised who were known to take at least one dose of treatment and who provided any follow-up data for one or more key efficacy variables; in turn, ITT patients are allocated to treatments actually received.*

From this definition it is clear that clinical arguments were deemed to outweigh statistical ones in some cases. In their view, the clinical objective of a study is to determine how a subject responds to a specific treatment and, as such, outcome has no meaning unless at least one dose is administered and the subject is analysed according to the treatment received. (This view may be considered somewhat atypical in statistical circles.) Similarly, subjects with no post baseline do not add to our knowledge of the treatment (apart from estimation of withdrawal) and attempts to include these subjects in the analysis must necessarily be based on strong and un-testable assumptions. Accordingly, Gillings and Koch described the function of such subjects in an analysis as *unclear and debatable.* Gillings and Koch also went a step further and attempted to quantify the acceptable proportions of excluded subjects from a modified ITT population as no more than 5% of

all randomised subjects, with the additional qualifier that such exclusions should be verified as independent of treatment and outcome. They also suggested that a PP population should include at least 80% of subjects from the modified ITT population to be credible. (Note that for the analysis of Safety, the situation is much clearer (ICH E9, 1998) with regulatory expectation being that the Safety population will only include subjects who have received at least one dose of randomised treatment.)

However, such practical definitions of ITT and of the FS, and specifically the exclusion of subjects who do not have any on-treatment data, have recently proved contentious from a regulatory standpoint. Although ICH E9 states that in some circumstances the exclusion of these subjects may be reasonable, it includes the caveat that the potential for bias must be carefully considered in each case. (It also makes a similar statement regarding subjects who do not take at least one dose of randomised treatment.) For instance, Phillips and Haudiquet (2003) found that their ITT modification of excluding subjects with no post baseline data was acceptable to most - but not all - European regulatory authorities, when seeking regulatory approval of their pain relief treatment. Now, in response to this finding, Brown (2003) - a statistician from the MHRA - expressed concern with the Phillips and Haudiquet practical definition, despite acknowledging that it is common practice in drug development. Understandably, Brown's main concern related to cases were the proportion of subjects with no on-treatment data was not small and where this could be related to randomised treatment - as was indeed the case with the data presented by Phillips and Haudiquet. In this respect, Brown's comments are actually consistent with ICH E9 and his stance may be viewed as a simple re-enforcement of the basic principles surrounding data exclusion. As an alternative to the exclusion of subjects with no on-treatment data, Brown states a general preference for the imputation of missing data in order to produce a conservative estimate of the treatment effect that includes all randomised subjects. He

suggests that the carrying forward of baseline data (LOCF) may in some cases be the appropriate method, although it is not clear that this would necessarily be appropriate here. For instance, it is self evident that if similar proportions of subjects in each treatment group had no on-treatment data, then in terms of change from baseline, these subjects represent a cohort with both a treatment difference and associated variance of zero. In this respect, if such subjects were included in an analysis population, one would expect a non-null treatment difference to be underestimated and the corresponding variability of the estimate to decrease. For superiority trials the diminished treatment difference would be conservative although the reduced variability would actually be anti-conservative from a hypothesis testing perspective. (Clearly for equivalence studies the approach would be anti-conservative overall). However when the proportions of subjects with no post-baseline data differ between the treatment groups it is not clear that substituting forward baseline data would necessarily constitute a conservative approach.

What this exchange of views between statisticians emphasises is that the greatest practical challenge faced when attempting to implement the ITT principle in its purist form - from both clinical and statistical perspectives - is the handling of subjects with missing data. As suggested earlier, whereas misclassification associated with procedural non-compliance could be considered in some cases simply to represent the estimation of the treatment difference in practice, missing data are different proposition altogether, since by definition these do not contribute to the estimation procedure without some form of imputation. However, as Koch, Davis and Anderson (1998) state, *the principal dilemma for missing data is that there is no 'clearly correct' method for managing it.*

In broad terms, two options are available - analyse only the non-missing data (that is, ignore the missing data), or alternatively, impute data and observe strict adherence to the

ITT principle. ICH E9 makes the general statement that *imputation techniques, ranging from the carrying forward of the last observation to the use of complex mathematical models, may also be used in an attempt to compensate for missing data.* It also highlights the simple dichotomization, success or failure, which can be applied to ensure that all subjects are included in the analysis population. However the coverage of missing data in ICH E9 is limited and more recently, the CPMP has produced a specific Points to Consider (PtC) on missing data (CPMP/EWP/1776/99, 2001). This document states that *the statistical analysis of a clinical trial generally requires the imputation of values to those data that have not been recorded* and emphasises that *bias is the most important concern from missing data* – in terms of estimation, baseline comparability and representativeness. However it includes no reference to the minimisation of bias, preferring instead to refer to the need for conservative methods that do not favour the study objective – essentially re-enforcing the message that superiority and equivalence (or non-inferiority) trials are viewed differently by regulatory agencies with regard to *the types of bias that affect interpretation.* Pre-specification of the approach to handling missing data in the study protocol is a key component of the PtC and the expectation is that selected methods should be *optimal* - since *different approaches may lead to different results* and that the methods employed may actually introduce bias. It also recommends that the degree of missing data is predicted at the design stage and that a statement is made regarding the acceptable level of missing data. The key role of sensitivity analysis is emphasised – an approach whereby the influence of various methods of handling missing data are investigated by comparing the different results obtained – and if the results obtained are similar then the findings can be considered robust. In this respect, in contrast to the comparison of different populations, where one could argue that different parameters are being estimated, a sensitivity analysis for missing data is simply a case of comparing different estimates of the same parameter.

Regarding methods for handling missing data, the PtC elaborates on some of the highlighted approaches - although the approach of simply using observed data when some are missing is regarded as unacceptable for a primary analysis. In this case the PtC raises conflicting concerns - the potential for missing data to have been more extreme leading to variance underestimation versus the reduced power due to fewer observations. Regarding imputation, the popular LOCF approach is viewed positively if used conservatively - for instance, if the condition is known to improve with time and the test treatment has a greater proportion of withdrawals. It cautions against using LOCF for deteriorating conditions such as Alzheimer's disease - although conceivably the approach would be conservative if there were fewer withdrawals with test treatment compared with reference. In general, the acceptability of LOCF appears dependent upon the expected time course of the disease together with an assessment of differential withdrawal - including an evaluation of the specific reasons for withdrawal. However carrying forward baseline data is not specifically addressed and the suggested tactic of *post hoc* evaluation of unblinded data does increase the risk of introducing bias through the adoption of driven analysis conventions. (Note that according to Gillings and Koch (1991), LOCF is generally an *even handed* approach and acceptable if <10% data are missing.) Other imputation methods such as estimating single or multiple values from other study participants is discussed relatively favourably with the cautionary note that single imputation methods tend to reduce variability. Interestingly no direction is given as to which variables the imputation should be based on. For instance, given the ICH E9 requirement for a study to be analysed as it has been designed, it would also be appropriate to "impute as designed" and include treatment and stratum in the algorithm as a minimum. There is also general advice to minimise the occurrence of missing data by making every effort to collect follow up data on subjects who withdraw or who violate the protocol.

The PtC mentions the extreme imputation method whereby all subjects with missing data are assigned the worst outcome in the in the reference treatment group but the best outcome in the control group (and *vice versa*). However it has been shown that even with only modest amounts of missing data, these two approaches would be expected *a priori* to produce inconsistent results and as a sensitivity analysis is therefore relatively uninformative unless few subjects have missing data, in which cases all methods generally produce similar conclusion (Unnebrink and Windeler, 1999).

An alternative less extreme imputation method that is commonly employed for binary outcomes is to assign all subjects with missing outcomes as default failures. (A similar but less common approach is to assign all subjects with missing data as default successes.) However although this method is simple and widespread, the consequences in relation to bias can be varied. It can be shown how bias can be introduced through the route of differential missingness by adapting the misclassification model (3) from before, where it was assumed that $H_0 : \pi_t = \pi_r = \pi$. Now, if subjects with missing outcomes are systematically assigned as default failures then the false positive rate for this imputation procedure is zero - that is, $\varphi_t = \varphi_r = 0$. In this case,

$$E(p_t' - p_r') = \pi(\theta_r - \theta_t)$$

and in the presence of different false negative rates, bias increases as π approaches one. Similarly, if subjects with missing outcomes are systematically assigned as default successes then $\theta_t = \theta_r = 0$, and

$$E(p_t' - p_r') = (1 - \pi)(\varphi_t - \varphi_r)$$

In this case, in the presence of different false positive rates, bias now increases as π approaches zero. It follows that unless the probability of being missing is identical in both treatment groups, the estimated treatment difference can be biased in either direction

depending on the pattern of missingness. Furthermore, the customary method of assigning default failure status may actually maximise the potential for bias in those disease areas such as anti-infectives where π approaches one and is similar in each treatment group.

From a regulatory perspective, it is clear that the challenges faced when addressing analysis populations and missing data are tightly interwoven. Moreover, in both cases the phrase conservative is referred to in regulatory guidance and it is interesting to consider this aspect further. The Oxford dictionary (1993) definition of conservative includes the phrases: *characterized by caution, moderation*; *(of views, taste, etc.) avoiding extremes*; *of an estimate etc.: purposely low*. Certainly in the general sense one would expect regulatory agencies to adopt a cautious approach to drug approval in their role as public watchdog, while at the same time avoiding extreme views. However the phrase *purposely low* in relation to an estimate is interesting in that it implies a deliberate attempt to control direction.

Now, the focus of ICH E9 is statistical principles, and it states that *many of the principles ... deal with minimizing bias ... and maximising precision*. At the design stage it is clear that the topics covered by ICH E9 are aimed at minimising bias - for instance randomisation and blinding. The principle of pre-specification (directed at both the protocol and the statistical analysis plan) is aimed at ensuring objectivity and therefore minimises the chances of introducing bias at the reporting stage. Principles directed at study conduct (such as interim analyses and use of independent data monitoring committees) also minimise bias through pre-specification, maintenance of blinding etc. At the analysis and reporting stage however the underlying principle of bias minimisation is less clear despite the requirement for pre-specification.

Some areas are uncontroversial in this respect. The section in ICH E9 on handling outliers is explicit in the requirement to select a procedure *such as not to favour any treatment group a priori*. Subgroups should be pre-specified or considered exploratory only, while data transformation must also be pre-specified. Covariate adjustment is primarily aimed at increasing efficiency and for standard linear models this equates to maximising the precision of estimates – although if there is imbalance at baseline with regard to an influential covariate then covariate adjustment may also be considered to minimise the bias of the estimate of the treatment difference. (For other linear models (generalised) - including the logistic and proportional hazards formulations - covariate adjustment tends to increase the estimate of the treatment difference and decrease precision, although the overall impact is increased efficiency.) Again pre-specification is key to minimising the introduction of bias and provides protection from an approach that selects the most favourable model following treatment unblinding. Furthermore, the restriction of covariates to those recorded at randomisation (that is, pre-treatment) minimises the introduction of bias that could result from including covariates with values related to treatment. However in the related areas of analysis sets and missing data, this situation in relation to the minimisation of bias is less clear.

The section pertaining to analysis sets in ICH E9 initially re-enforces the principle of bias minimisation. However, as described previously in Section 2.3.1, the full analysis set is envisaged *to avoid over-optimistic estimates of efficacy* and as such represents a conservative strategy for superiority trials, whereas for equivalence or non-inferiority trials it is deemed anti-conservative. The safety set could also be viewed as conservative in the sense that ICH E9 points to the exclusion of subjects who did received at least one dose of the investigational drug. In this respect the denominator for adverse event estimation is potentially reduced leading to higher event incidences. Although uncontroversial, it points

to the direction of any potential bias being more important than the potential bias introduced as a result of subject exclusion. In terms of missing values, ICH E9 emphasises the importance of pre-specification in terms of procedures/data conventions for handling missing data and also the requirement for sensitivity analyses. However the earlier illustration taken from the PtC on missing data which states: *in depression, where the condition is expected to improve spontaneously over time, this method [LOCF] might be considered conservative if patients in the experimental group tend to withdraw earlier and more frequently due to safety reasons*, represents a direct statement to the effect that the term conservative would represent a scenario where the direction of the bias naturally tended towards shrinkage of the treatment effect. The PtC later states in relation to best or worst case imputation that this approach *may be considered, provided it is applied conservatively*. It continues: *These techniques may be useful to assess a lower bound of efficacy as a demonstration of robustness*. On the other hand, the PtC discusses other imputation methods - such as maximum-likelihood and multiple imputation - and mixed models – neither of which would necessarily be conservative. Interestingly in the context of the pre-specification of missing value procedures, the PtC stresses that it is *of particular importance to ensure that the selected method is a conservative approach and does not favour the study's working hypothesis (intentionally or unintentionally)*. That is, avoiding underestimation with non-inferiority hypotheses whilst avoiding overestimation with superiority hypotheses. On balance therefore one could argue that the regulatory requirements in relation to the related topics of analysis sets and missing values more accurately represent an attempt to control the direction of the bias rather than to minimise it *per se*.

Indeed it is for this very reason that the *Points to Consider on switching between superiority and non-inferiority* (2000) is so compelling since it brings into conflict two sets

of diametrically opposed approaches – one aimed at producing a conservative estimate of the treatment difference for superiority, the other aimed at producing a conservative estimate for non-inferiority. Given that the switching strategy specified in the PtC is aimed at drawing the correct conclusion from the values of the estimated confidence limits – and once the data are observed it is only the conclusion that may change and not the confidence limits themselves - it is clear that the juxtaposed statistical conventions relating to the conservative analysis of both superiority and non-inferiority hypotheses must be reconciled. Perhaps this juxtaposition affords the opportunity to select a more neutral approach actually aimed at minimising potential bias – in this context the phrase *optimal* from the PtC on missing data is appealing.

It should not be forgotten however that the role of sensitivity analyses is integral to regulatory review strategy and is particularly important in relation to analysis sets and missing values. The complementary role of sensitivity analysis ensures that the results from different analysis and the application of various data conventions are considered in their totality. In this respect attempts to control the direction of the bias through conservativeness actually informs the decision making process. It is also clear that the regulatory guidelines encourage good study conduct to minimise the impact of issues, such as missing values and protocol violations, on statistical analyses.

In summary, it is s clear is that when aiming for a purist form of ITT, the management of potential bias – regardless of whether the aim is minimisation, directional control or optimisation – depends upon the observed pattern of missing data and the underlying assumptions that are ultimately impossible to validate. Therefore, although ITT may be considered unbiased in terms of the comparability of the treatment groups at randomisation, the practical challenge of implementation often means that the introduction

of other types of bias cannot be ruled out. Nevertheless, ITT and the underlying principles behind it represent the most coherent approach to clinical trial reporting and perhaps Fisher *et al* (1990) provide the most insightful conclusion on the choice of analysis populations when they state: *We feel that the intent-to-treat analysis may not always be the best analysis but when it is not this usually indicates that (1) the experiment was designed or run sub-optimally and (2) the results are even more debatable.*

The next section of this chapter, the challenge of sub-setting data progresses to subgroups. Regulatory considerations relating to subgroups will be discussed and in particular issues surrounding multiplicity.

2.4 SUBGROUPS

2.4.1 General considerations

The primary purpose of subgroup analyses is to investigate whether treatment differences are consistent within a study for different levels of a factor (or combinations of factors), and according to Pocock *et al* (2002), these baseline factors should be selected on the basis of *scientific and ethical obligation.* The reference to baseline factors is important, since subgroups that are based on treatment independent factors - including those recorded at baseline - will generate an unbiased distribution of subjects to treatment groups. Accordingly, over all randomisations, the treatment groups will be balanced for all remaining factors and covariates, and standard tests of statistical significance remain valid within each subgroup.

Subgroups are generally defined by a single factor – gender for instance, with the subgroups male and female – although subgroups may also be defined by a combination of distinct factors - such as gender and age (≥65 years versus <65 years), say. Such

combinations may also be used create a hierarchy of subgroups. For instance Matcham (2003) describes the area of anaemia in renal failure whereby a regulatory agency not only requested subgroup analyses by route of administration (intravenous versus subcutaneous) but also separate subgroup analyses of subjects who received subcutaneous treatment by mode of dialysis (haemodialysis versus peritoneal dialysis).

Pocock *et al* (2002) identified several issues when undertaking subgroup analyses. For instance, despite most studies being inadequately powered to detect treatment differences within subgroups, multiplicity concerns remain. This is due to the almost unlimited number of subgroup analyses that could be undertaken in a study, and the resulting vulnerability to *post hoc* selection of the most appealing. For example, if ten independent subgroup analyses were undertaken at the 5% level of significance, then under the null hypothesis of no treatment difference, the chance of observing at least one statistically significant result would be $(1-0.95^{10}) = 0.40$. In their view, the most appropriate approach to determining consistency of treatment effect across subgroups is through investigation of the treatment by subgroup interaction - that is, the difference between subgroups in terms of the treatment difference. Although such tests of interaction are considered as having low power, Pocock *et al* view this property positively stating that *interaction tests recognize the limited extent of data available for subgroup analysis, and are the most effective tool in inhibiting false or premature claims of subgroup findings.* The final issue identified by Pocock *et al* is the degree to which subgroup analyses should influence the study conclusions.

Bennett (1993) also recommends the use of tests of interaction. He suggests a hierarchy of inference whereby one first estimates the overall treatment effect (main effect), then investigates the selected subgroup (subgroup effect). The next step is to determine whether

the magnitude differs between subgroups (quantitative interaction) and then whether the direction of effect is different (qualitative interaction). In his view inferences regarding the main effect and qualitative interaction have important implications in *formulating a general policy about treatment*, whereas subgroup and quantitative interactions are important to dose selection and understanding the biology of the disease.

Although both authors stress the importance of interaction analyses to compare subgroup differences, neither highlights the influence exerted by the scale of measurement when considering quantitative interactions. That is, a quantitative interaction observed using one scale of measurement can quickly disappear once transformed and *vice versa* (Hand, 1994). This point is illustrated in Table 2.I. In this case the odds ratio of 3 (Test/Reference) is identical for each subgroup (as defined by Factor F) whereas the difference in response percentages suggests an interaction of 10 percentage points. (Note that some (Nelder, 1994) would actually view the analysis based on the difference in proportions as inappropriate since unlike the logit transformation, where parameter values are in the real plane $(-\infty,+\infty)$, the difference in proportions is bounded in the unit square $(-1,+1)$.) Gail and Simon (1985) similarly describe how an interaction can disappear through the logarithmic transformation of continuous data. Indeed Gail and Simon state that *because there is usually no self-evidently appropriate scale of response measurement, quantitative interactions are to be expected, but they may not be important clinically.* This is in agreement with Peto (1982) who expects to observe quantitative interactions when factors, on which the subgroups are based, are known to influence outcome. The qualifier referring to factors that affect outcome is important since it is the different absolute effects of each level of the factor that affords the opportunity for the relative treatment effect to differ between subgroups on some selected measurement scale. (Indeed this is the case in

Table 2.I.) Accordingly, if there is no factor effect, then the scope for observing differential treatment effects simply through scale modification is markedly reduced.

Table 2.I. An example of the influence of the scale of measurement on the interpretation of an interaction

	Factor F=1	Factor F=2
Test treatment	90% (90/100) *(Odds=9)*	75% (75/100) *(Odds=3)*
Reference treatment	75% 75/100 *(Odds=3)*	50% 50/100 *(Odds=1)*
% Difference Odds ratio	15% 3	25% 3

One important feature of subgroup comparisons that is often not fully appreciated is that subjects' characteristics are not assigned at random (Senn, 1997). In this respect, a treatment by subgroup interaction may actually reflect differential sampling of subjects. For instance, a trial may exclude women of childbearing potential, and in this case the women randomised into the study may be older, on average, when compared with the males subjects. (Note that although the experimenter has no control over the assignment of individual characteristics such as gender, the actual genetic assignment [in terms of chromosomes] at the time of conception may actually be considered to be random.) Now, if the true treatment difference is related to age but not gender then an apparent gender difference may be observed due to the confounding. In this respect, although random allocation of the treatment assignment delivers causality to the overall treatment comparison – either conditionally or unconditionally - this does not extend to the determination of causality for the comparison between subgroups with regard to the treatment difference. Figure 1.1 (Chapter One) earlier illustrated a possible hierarchy of subgroups for the female subgroup where the effects of specific treatments may differ due to physiological changes or drug interactions. Other examples of potential confounding

factors for a number of common subgroups are gender (weight, body fat and hormones), race (weight and diet) and age (hepatic and renal function, and concurrent treatment)

Another related feature of subgroup analyses is that sometimes the categorisation of data into unique subgroups appears somewhat arbitrary – particularly if based on defined thresholds for a continuous variable, such as age. Furthermore the number of subgroups formed from a factor may also be arbitrary. In this respect, decisions regarding the actual number of subgroups and the associated thresholds have the potential to generate results of a purely invented nature since power is related to both the true treatment effect and the sample size. Pocock (1983) provides an example of such arbitrariness when metoprolol was compared with placebo with regard to the percentage of deaths in the treatment of acute myocardial infarction (Hjalmarson *et al*, 1981). Overall there was a statistically significant difference (odds ratio= 0.62) between the treatments (χ^2 test, p=0.023) although not surprisingly it was noted that the death rate increased with age regardless of randomised treatment. The treatments were compared within three subgroups dependent on age (40-64 years, 65-69 years and 70-74 years) to investigate consistency of effect and the re-calculated data are presented in Table 2.II.

Table 2.II. Percentage of deaths in treatment of acute myocardial infarction by age

Age	Placebo		Metoprolol		odds ratio (M:P)	χ^2 test p-value
40-64	26/453	5.7%	21/464	4.5%	0.78	0.40
65-69	25/174	14.4%	11/65	6.7%	0.43	0.021
70-74	11/70	15.7%	8/69	11.6%	0.70	0.48

Note: p-values re-calculated using StatXact software with odds ratios added for clarity

The largest treatment difference was in the middle (65-69 years) subgroup that singly achieved statistical significance. However, Pocock showed that by combining subgroups, apparently conflicting results and conclusions could be found. For instance the p-value for the combined 40-69 years subgroup is 0.030 while for the combined 65-74 years subgroup it is 0.023. As such, for the first combination the results show a treatment effect is exclusive to younger subjects while the second combination shows the opposite – that is, the treatment effect is exclusive to older subjects. (Note that Hjalmarson *et al* were careful not to over-interpret these data and did not claim a differential treatment effect based on age.) Pocock notes that if an interaction test had been performed then it would have been non significant. Indeed for completeness, Zelen's interaction test from StatXact (Cytel, 1999) gives p=0.48 while the estimated common odds ratio adjusted for the three age categories is 0.63 with an exact p-value of 0.034 – almost identical to the unadjusted result.

In the next sub-section, the current regulatory guidance in relation to subgroups will be explored while in the following two sub-sections, the observation of inconsistent results across subgroups and the issue of multiplicity will be discussed in detail.

2.4.2 Regulatory considerations

In recent times, regulatory authorities have begun to recognise the importance of presenting outcome data for particular subgroups of subjects. Indeed the ICH E3 guideline *Structure and content of clinical study reports* (CPMP/ICH/137/95, 1995) identifies a specific section - *Examination of subgroups* - in the ICH report template where subgroups should be discussed. ICH E3 gives some examples of *important demographic and baseline value-defined subgroups* that should be addressed. These are *age, sex, race, severity or prognostic group,* and *history of prior treatment with a drug of the same class.* Furthermore the expectation is that an explanation would be provided in the report if these

subgroup summaries were not presented. It also cautions *that these analyses are not intended to "salvage" an otherwise non-supportive study*. Subgroup analyses receive limited attention in the statistical guideline ICH E9. ICH E9 emphasises the need to pre-plan subgroup or interaction analyses for factors that are of specific interest, but also highlights that most analyses of this kind *should be interpreted cautiously* and are of an exploratory nature. It recommends interaction analyses, *complemented by additional exploratory analysis within relevant subgroups of subjects*. It also cautions that in relation to drug approval, conclusions *based solely on exploratory subgroup analyses are unlikely to be accepted*. However in 2001, the CPMP issued the *Points to Consider on multiplicity issues in clinical trials* (CPMP/EWP/908/99, 2001) that expands the ICH E9 guidance in relation to subgroups in Europe. This document addresses two specific areas in more detail - the acceptable circumstances for claiming a subgroup effect and the potential license restriction to specific subgroups. Now, the need to provide compelling evidence for a specific subgroup claim can be interpreted as a regulatory requirement to control the α error or false positive rate. In this respect, reliable conclusions leading to regulatory acceptance are achieved through the pre-specification of subgroups, with formal consideration given at the design stage to stratification and power. Furthermore the expectation is that the overall treatment difference would already have achieved statistical significance, suggesting a step-down procedure to control the α error in the subgroups. The alternate side of the coin, is the undesirable restriction of a licence to specific subgroups, and in this respect the requirement can be viewed as the need to demonstrate consistency or uniformity of effect across subgroups of known importance. In this case, the control of the β error or false negative rate is the issue of regulatory concern, and although not stated explicitly, it is the interaction between treatment and subgroup that is of interest in this regard. The document describes how strong heterogeneity of effect between subgroups (that is, a qualitative interaction) might lead to licence restriction. This would be the case

when such effect *can reasonably be assumed but cannot be sufficiently evaluated* with the observed data, but also when an unexpected effect is observed and cannot be explained - in both instances, further clinical data would be required to lift the licence restriction. (The phrases *reasonably assumed* and *sufficiently evaluated* are in fact quite interesting and perhaps signify a well hidden regulatory desire for pre-specification of expected effect size, with evaluation of the observed data against this expectation - a point re-visited in Chapter Four.) The document also adds to the list of potentially important factors given in ICH E3, with the inclusion of geographic region, renal impairment, and drug absorption and metabolism. One further relevant guideline in relation to subgroups is the CPMP's *Points to Consider on adjustment for baseline covariates* (CPMP/EWP/2863/99, 2003). This document refers to subgroups and recommends the pre-identification of subgroups where expectation exits that *substantial interactions* are likely to be present. In this case, it recommends that either each individual subgroup should be adequately powered to detect the required treatment difference or the study should be restricted to just one subgroup through the application of specific inclusion criteria. The PtC references ICH E9, and re-enforces the view that the investigation of treatment by subgroup interactions is regarded as exploratory due to limitations of power, and that non significant findings do not constitute evidence of no interaction. However the document cautions that the primary model excluding the interaction term could be *invalidated* in cases where the interaction is *particularly strong or even qualitative* and *the results of trial could become inconclusive.*

2.4.3 The observation of inconsistent results across subgroups

In this sub-section it will be shown how likely it is in practice that inconsistent results will be obtained across subgroups. It will also be shown how the tactic of using subgroup analyses to demonstrate consistency of effect is usually under-powered and arguably flawed. However since in practice subgroup analyses are often requested by regulatory

agencies for both superiority and non-inferiority studies, the tactic of using tests of directional advantage will be discussed as a means of providing adequate overall power.

Peto (1982) gives a simple illustration of how inconsistent results are easily obtained when comparing two subgroups with regard to the difference in a continuous outcome between two treatments - test (t) and reference (r), say. Let the variance of the overall treatment difference, $Var(\bar{x}_t - \bar{x}_r)$, be v, in some arbitrary units, from which it follows that the variance of the treatment difference, $Var(\bar{x}_{tj} - \bar{x}_{rj})$, within each of two subgroups (j=S, S', where S' is the complement of S) of equal size will be $2v$, while the variance of the interaction term, $Var(\hat{\Omega})$, will be $4v$. (Refer to Chapter Four, Sub–section 4.2.2 for further detail.) Furthermore, by definition, there is around a 1 in 3 chance that an observed interaction will be $> 2\sqrt{v}$ even though no interaction exits (that is, Ω=0) since 68% of the Normal distribution is contained within the region $0 \pm \sqrt{4v}$. Now, it follows that if the observed overall difference between the treatments is $2\sqrt{v}$ - with a resulting p-value of borderline significance (that is, 0.05) - then there is also a 1 in 3 chance that the observed treatment difference in one subgroup (S, say) will be $> 3\sqrt{v}$ and $< \sqrt{v}$ in the complement (S'). (This can be shown by solving the simultaneous equations $(d_s - d'_s = 2\sqrt{v})$ and $\left(\frac{d_s + d'_s}{2}\right) = 2\sqrt{v}$, where d_s and d'_s are the treatment differences in the subgroups S and S' respectively.) In this case, the treatment difference in subgroup S will be statistically significant ($p<0.05$), while in S' the difference will be non significant. That is, $\frac{3\sqrt{v}}{\sqrt{2v}} > 1.96$ and $\frac{\sqrt{v}}{\sqrt{2v}} < 1.96$ in S and S' respectively. This illustrates that even when no treatment by subgroup interaction exists, if the overall treatment difference is of borderline

significance, then the chance of observing statistically inconsistent differences in two equally sized subgroups is not insubstantial – that is, around 1 in 3.

Koch (see Koch and Gansky 1996, Koch 1996, Koch 1997) has undertaken some of the most incisive work in the area of subgroup analysis in relation to drug development and regulation. As a means of investigating the prospective power of subgroup analyses for a continuous outcome, Koch (1997) relates this to the significance level and power of the overall treatment comparison, and to the chosen significance level for the corresponding within subgroup comparison. In this case, Z_α and Z_β represent the 100(1-α) and 100(1-β) percentiles of the standard normal distribution and the additional subscript "s" refers to the subgroup comparison. (Note that for two-tailed significance tests, α is multiplied by two.) Now, Koch shows that the power (1 - β), for a subgroup analysis containing $n_s = f_s n$ $(0 < f_s \leq 1)$ subjects in each treatment group, where n is the original number of subjects in the treatment group, can be calculated as:

$$Z_{\beta s} = \left\{ \frac{\sqrt{f_s}\left(Z_\alpha + Z_\beta\right)\delta_s}{\delta} \right\} - Z_{\alpha s} \qquad (4)$$

where δ is the overall expected treatment difference and δ_s is the corresponding expected treatment difference in the subgroup.

Now, when $f_s = 0.5$, $\delta_s = \delta$ and $Z_\alpha = Z_{\alpha s}$, this reduces to

$$Z_{\beta s} = \frac{\left(1-\sqrt{2}\right)Z_\alpha + Z_\beta}{\sqrt{2}}$$

It is then straightforward to show that a trial planned with 80% power $(Z_\beta = 0.84)$ to detect an overall treatment difference of δ at the two-sided significance level of 5% $(Z_\alpha = 1.96)$ provides only 50% power $(Z_{\beta s} = 0.020 \approx 0)$ to detect an identical treatment

difference of δ in a subgroup containing half the subjects. Furthermore, it follows that the power to achieve significance simultaneously in two subgroups is just 25%. That is, 0.5^2. Accordingly, the probability of observing an overall inconsistent result – one subgroup significant, the complement subgroup not significant – is 50% while the probability of neither result being significant is 25%. Therefore the tactic that attempts to show consistency of effect through individual subgroup analyses is clearly under powered for the typical clinical trial.

(Note, that if the time to some event is the outcome of interest, then the number of events rather than the sample size *per se* determine power (Sleight, 2000). In these cases, subgroups expected to contain few events will have dramatically less power than those with many events, even though the numbers of subjects expected in the subgroups may be similar. Furthermore for binary outcomes, the expected variability is related to both the sample size and the expected response proportions such that subgroups with response proportions close to 0 or 1 will have the greatest power to detect differences in proportions.)

Now, as a means of accounting for these substantial reductions in power, Koch suggests that the significance level for each individual subgroup comparison could be actually increased - to an extent that results in sufficient power for both subgroup analyses simultaneously. For instance, he suggests that adopting a one-sided α_s=0.1 significance level for the subgroups when the overall comparison is planned to have 90% power gives 85% power for each of two subgroups with simultaneous power of around 72%.

As an indication of just how high α_s would need to be raised, if the aim were to plan to maintain power for the simultaneous subgroup analyses at the same level as the overall

comparison at the two-sided 5% significance level, then consider the case where $f_s = 0.5$ and $\delta_s = \delta$, and (4) reduces to

$$Z_{\alpha s} = \left\{ \frac{(Z_\alpha + Z_\beta)}{\sqrt{2}} \right\} - Z_{\beta s} \qquad (5)$$

Now, if the overall comparison were planned to provide 80% power, then each of two subgroups would require 89.4% power individually to provide 80% power simultaneously. As such, $Z_{\beta s} = 1.25$, $Z_\beta = 0.84$ and $Z_\alpha = 1.96$, in which case $Z_{\alpha s} = 0.73$, giving a two-sided α of 0.465 or a one-sided 0.233.

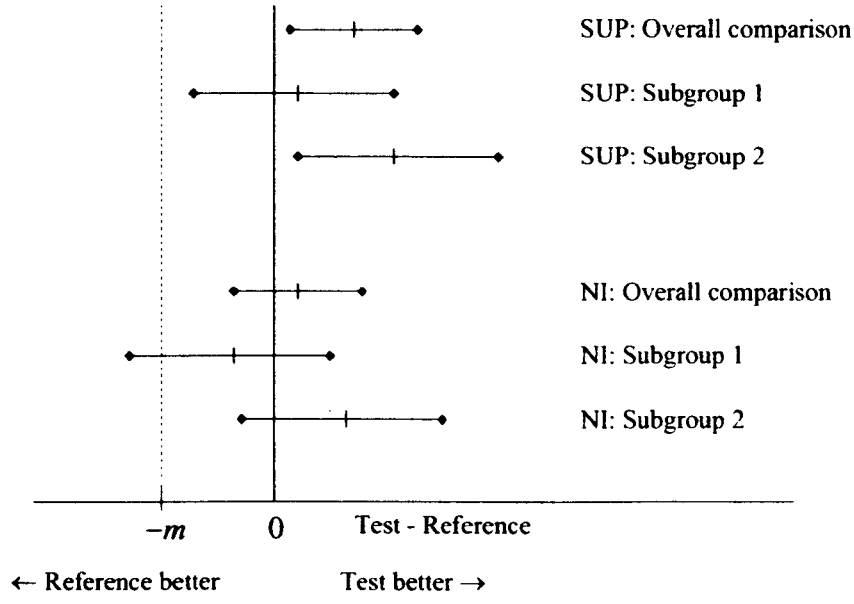
Now, the illustration is actually quite interesting in its relation to observations and comments from regulatory statisticians. For instance, it has been stated informally that regulatory statisticians at the MHRA look for consistency of the point estimates of the treatment difference when reviewing subgroup analyses, and in particular concerns are raised when these point estimates are on the wrong side – that is when the sign of the point estimate is reversed in a particular subgroup (one-sided test p>0.50). Now, this informal rule is broadly similar to an approach proposed by Koch. Koch suggested that to achieve adequate power - when aiming to demonstrate consistency of effect through the achievement of $p \leq \alpha_s$ in all subgroups for all factors investigated - then α_s may need to increase to as high as a one-sided 0.50. That is, adoption of a tactic of testing for directional advantage – although Koch's preference was actually not to exceed a one-sided 0.25. In this respect, Koch's point is that analysis tactics require adequate power (≥70%, say) so that failure to reject one hypothesis is interpretable as evidence against consistency of effect. Now, re-arranging (5) and setting $Z_{\alpha s}$ to zero for a one-sided test at the 0.50 level of significance, implies the power for a single subgroup would be $Z_{\beta s} = 1.98$ - that is, 97.6% power. Furthermore, for the simultaneous testing of more than one subgroup, such a tactic would provide 80.5% power for 9 subgroups (each contained 50% of subjects)

while 14 subgroup investigations would provide the maximum number to meet Koch's criterion of ≥70% power. This observation demonstrates that if the MHRA – and perhaps other regulatory agencies – are assessing consistency of effect through one-sided directional tests at the 0.50 level of significance, this approach will provide reasonable power so long as the number of factors is few, five say, and have only two levels that partition the subjects equally, producing in the region of 10 subgroups.

(Note also that there are some similarities here to the approach to the design and analysis of pragmatic trials proposed by Schwartz, Flamant and Lellouch (1980). They proposed that in the comparison between two active treatments, a study should be powered to control the so-called γ error – defined as the probability of reaching a conclusion with the wrong sign, that is, *the recommendation of an inferior treatment*. In this case, the treatment comparison essentially becomes a decision making challenge with a two-sided α set to 1 and β to 0. That is, the treatment that is numerically superior is simply selected.)

The directional advantage approach is also helpful in formulating how subgroup analyses could be interpreted in a non-inferiority study. (Recall that in such designs, a margin (*-m*, say) is pre-specified for the difference between the test and reference treatments such that non-inferiority can be claimed in cases were the 95% confidence interval for the treatment difference excludes values less than *-m*.) In this case the consistent approach would be to require that the point estimate of the treatment difference for each subgroup is >*-m*. Figure 2.1 illustrates separately the nature of a successful outcome when a directional advantage approach is applied to superiority or non-inferiority studies. In each case the subgroups satisfy the directional test criterion although the lower confidence limit is to the left of the margin (zero for superiority, *-m* for non-inferiority) for Subgroup 1.

Figure 2.1. The directional advantage approach for superiority and non-inferiority studies (mean treatment difference and 95% confidence interval).



In reference to non-inferiority trials, it has also been mooted that the MHRA expect to see the point estimates for subgroup analyses greater than zero rather than $-m$. Clearly if the treatments are truly equivalent then the probability of achieving this in each subgroup is 50%. Although when conducting non-inferiority studies it is often the case that test treatment is considered marginally better than reference (the basis of the MHRA's case), it is arguable that this approach would be ultra-conservative in many cases.

2.4.4 Multiplicity

According to Assmann et al (2000), of all the various multiplicity problems in clinical trials subgroup analysis remains the most overused and overinterpreted, while in the view of Senn (1997), subgroup analysis ought to be taken as an admission of a lack of confidence in the pooled estimate. Assmann et al also recognises that most subgroup

claims are prone to exaggerate the truth. Now, in relation to multiplicity, Koch (1996) makes an extremely important observation in relation to the reporting of subgroup analyses to regulatory authorities. In his view, since confirmation of homogeneity of effect is the prime reason for performing subgroup analyses, adjustment for multiplicity is generally not required due to the high probability of observing inconsistent results that leads not to a strengthening of the conclusion but to the inevitable dilution of it. Furthermore, according to Koch and Gansky (1996), the most appropriate tactic is to manage the type II error by aiming to use a significance level of 5% as much as possible whilst also maintaining the significance level collectively for all assessments at 5% level.

One approach to implement such a tactic is to use the closed test principle (Marcus *et al*, 1976) where the set of null hypotheses to be tested is required to be closed under intersection (Bauer 1991, Senn 1997). In this respect an individual null hypothesis can only be rejected at the α significance level if all other higher level hypotheses containing it are also rejected at the α significance level. Now, in terms of applying the closed test principle to subgroups, rejection of the null hypothesis that there is no overall treatment difference supports the claim that a treatment difference exists in at least one subgroup. Consequently, if a factor exists that defines two subgroups - S_1 and its complement S_2 - then the null hypothesis of no treatment effect can then be tested in each subgroup separately also using a significance level of α. Furthermore, if subgroup S_1 is further sub-divided by another two-level factor producing subgroups S_{11} and its complement S_{12}, then providing the null hypothesis of no treatment difference was rejected for subgroup S_1, then a significance level of α can also be applied to test the null hypothesis of no treatment difference in each of these subgroups within a subgroup.

In cases were a factor has more than two levels, further steps are required to adhere to the closed test principle. For instance, if there are three subgroups (S_1, S_2, S_3), then the overall test represents the intersection $S_1 \cap S_2 \cap S_3$ hypothesis. Now, if this overall hypothesis is rejected at the α level then the intermediate intersection null hypotheses, $S_1 \cap S_2$, $S_2 \cap S_3$ and $S_1 \cap S_3$, must first be considered prior to testing the null hypotheses in each of the three subgroups (S_1, S_2, S_3). In this respect, to be able to test the null hypothesis in subgroup S_1 at the α level, both intersection null hypotheses $S_1 \cap S_2$ and $S_1 \cap S_3$ containing S_1 must both be rejected first at the α level. With a four level factor an additional level (or hierarchy) of testing is added, and so on.

Koch and Gansky (1996) suggest an alternative hierarchical method to control the α level that avoids the intermediate step of the closed test procedure above in cases were a factor has greater than two levels. They describe how following rejection of the null hypothesis for all patients, subgroups S_1, S_2, S_3 etc. are tested in a pre-specified order (S_3, S_1, S_2 say) at the α level until such time that the null hypothesis is not rejected at the α level. Indeed with all methods of adjustment the important aspect is to identify the structure of decision making *a priori* such that an appropriate testing strategy can be put in place (Norwood, 1996).

Clearly pre-specification is also an important aspect of controlling the type I error even when analyses are regarded as exploratory in nature, although as Sleight (2000) highlights, pre-specification should include not only the factor of interest but should also extend to stating the expected result. This approach has clear advantages when attempting to explain the biological plausibility of the findings at the reporting stage.

In cases where factors have not been pre-specified, adjustment may be appropriate even in a regulatory setting. For instance, it is not uncommon for regulatory agencies to request additional subgroup analyses following initial review of the CTD. One simple approach to adjustment when more than one factor is used to create subgroups is to employ a Bonferroni correction (Holm, 1979). The Bonferroni correction controls the overall α error at or below a predefined level, typically 5%, by testing all k hypotheses at the α/k level of significance. This method is extremely simple to apply, and indeed if the number of subgroups analysed has been specified, reviewers may make this correction manually themselves. (Note that α/k is actually an approximation, and the exact nominal significance level is given by $1-\exp\left\{\frac{1}{k}\ln(1-\alpha_o)\right\}$, where α_o is the overall α error, usually 5%.) However the Bonferroni correction is generally conservative in practice (although perhaps less so than for some other areas of multiplicity since subgroups are typically independent), and later adaptations of the approach control the overall α error whilst providing great power. For instance, Holm (1979) uses sequential rejection whereby hypotheses are rejected one at a time until failure to meet the rejection criteria occurs, at which point testing terminates. In this case the n ordered p-values (smallest p-value first) are tested at the α/k, $\alpha/(k\text{-}1)$, ... α levels of significance.

Another area where adjustment for multiplicity may be appropriate is in the evaluation of interaction tests. Unlike individual subgroups, the issue is not a case of observing inconsistent results for different levels of factor. Rather, tests of interaction are often judged in isolation and if numerous tests are performed then some adjustment may be prudent using the approach of Holm, say – particularly if factors are chosen as a matter of routine rather than biological plausibility.

In 1983, Pocock observed that *Unfortunately, with qualitative data interaction tests are complicated to perform and hence one may prefer to use them only when there is a strong indication that a genuine interaction might exist.* This point made in his influential book is revealing since it perhaps explains why subgroup analyses have to some extent prevailed when more appropriate methods for determining subgroup consistency are available (Pocock *et al*, 2002). That is, subgroup analyses are considered straightforward to perform since they represent a simple repeat of the overall statistical analysis in fewer subjects. Assmann *et al* (2000) reviewed 50 consecutive clinical trial reports published in four of the most respected medical journals (*NEJM, The Lancet, JAMA* and *BMJ*) during 1997. Assmann *et al* found that 35 reports (70%) included subgroup analyses, of which 7 were descriptive summaries, 13 provided within subgroup significance testing but less than a half (15) included tests of treatment by subgroup interaction. (Interestingly, in the same investigation, Assmann *et al* found that the fundamental building block for rational subgroup analysis - that is, pre-specification - receives less careful attention than one would like.) Of course, in the 21st Century, software advances have made interactions tests much more straightforward to undertake, and perhaps the true challenge now is to persuade researchers that this is the most appropriate method to assessing consistency of effect.

To complete this Chapter, some common examples of biased sub-populations and subgroups will be presented to illustrate the potential pitfalls of forming subsets that are simply intuitively appealing.

2.5 INHERENTLY BIASED SUB-POPULATIONS AND SUBGROUPS: SOME EXAMPLES

In this section, a selection of inherently biased sub-populations and subgroups are presented that are commonly reported.

2.5.1 Randomised cohort designs

Phase I drug development provides an interesting example where treatment groups are sometimes combined during the statistical analysis in a way that violates the randomisation principle. In a randomised cohort design, an initial cohort of subjects (perhaps eight subjects) is selected at a single centre, and these subjects are randomised to test treatment or placebo according to an unequal ratio (3:1, say). A low dose of test treatment is selected and the aim is to demonstrate that this dose is safe, and that an incrementally higher dose can be used in a second cohort containing different subjects to the first (that is, a cohort independent of the first). The placebo subjects are included to provide some limited control data to eliminate gross experimental effects at the study centre that could lead to premature termination of dose escalation procedure due to toxicity concerns - for instance, a virus causing diarrhoea or flu-like symptoms in all subjects in the cohort. A limit (say four or five) is placed on the number of cohorts and the study is terminated when either the maximum tolerated dose is observed or the final planned cohort is complete. In this respect each cohort can be regarded as a separate subgroup determined by sequential entry into the study – that is, determined by a time cut-off in each case. Now, at the reporting stage it is not uncommon to find that the safety profiles of each dose of test drug are compared not only with each other but also with a composite placebo group that includes all subjects from all cohorts. In fact both types of comparison violate the randomisation principle since this only holds within a cohort. Such pseudo dose response analyses are particularly vulnerable to seasonal effects and causality is impossible to determine. Apart from the valid within cohort analyses, one valid analysis that includes all cohorts is to fit a model that includes both cohort and treatment group (active versus placebo) - although note that this simply compares test treatment, across all dose groups, with placebo. Interestingly these randomised cohort designs have sometimes been employed for proof of

concept efficacy studies (Phase II). Despite the flawed nature of the dose response evaluation these studies have been conducted in substantial numbers of subjects in the multi-centre setting.

2.5.2 Duration of response

This category is a very fruitful area for the reporting of biased sub-populations/subgroups and examples from three different therapeutic areas – migraine, gastro-intestinal disease and oncology – will be discussed.

Migraine is a therapeutic area where interest is centred not only on whether an initial response to treatment is observed but also on whether the response is maintained. In these studies it is standard practice to require subjects to classify an emerging headache severity on a four-point scale as none, mild, moderate or severe. Once a headache has reached grade moderate or severe severity then it can be treated with randomised treatment and subjects are then required to evaluate the resulting severity in diary cards for a period 48-hours post treatment. A response to treatment is defined as a change in headache severity from severe or moderate to mild or none, and a positive response to treatment is maintained until such time that the headache returns or increases in severity from mild to moderate or severe (CPMP/EWP/788/01, 2002). For the primary analysis, treatments are typically compared with regard to the proportion of subjects with a response at fixed time points - 2 hours post-treatment, for instance. However maintenance of response is also important and typically the treatments have been compared with regard to the proportion of responders at 2 hours who relapse within 48 hours of treatment. However since response is not independent of treatment, this sub-population will be biased and so will be the estimated treatment difference. An alternative analysis is to consider treatment failure as the outcome and to retain all treated subjects – even those subjects that did not have a

response at 2 hours. This approach preserves the benefits of the randomisation, and in this case treatment failure is defined as the earliest of either no response to randomised treatment, the use of rescue medication or relapse during the 48-hour post treatment period. (However note that since subjects are randomised based on previous migraine frequency, some subjects will complete the study without treating a migraine attack and these subjects are excluded from the analysis population without the introduction of bias.)

The second example is from the study of gastric or duodenal ulcers. In the initial acute phase of these studies, high dose randomised treatment is usually assigned with the aim of healing the subjects' ulcers within 8 weeks of randomisation, say. Once healed, subjects then enter a long-term (perhaps 12 months) maintenance phase with a lower dose of the original randomised treatment. In a similar manner to migraine, it was common practice to analyse these studies as if they were two independent trials - an acute healing phase and a maintenance phase investigating relapse that excluded subjects who were not healed within the acute treatment phase. However this analysis is biased, as acute healing is not independent of treatment and hence unhealed exclusions are treatment related. An alternative appropriate analysis, that preserves the benefits of the randomisation, is to regard all unhealed subjects as treatment failures in the maintenance phase, with the time to relapse equal to zero. An alternative design has been used where all subjects receive test treatment in the acute phase but are randomised to treatment on commencement of the maintenance phase. Although this design is unbiased in terms of the randomisation, it can be criticised for selecting test treatment responders for the randomised maintenance treatment phase and in this respect it can also be considered biased. Another alternative design is to randomise subjects to acute treatment and then re-randomise healed subjects to maintenance treatment using acute treatment as a stratification variable. This is perhaps

the most even-handed design for both treatments and in this case the analysis of the maintenance phase includes acute treatment as a factor in the model.

The final example is from oncology where subjects are assigned to a number of cycles of randomised treatment. In simple terms, if a subject shows a defined reduction in tumour volume (partial response PR) or the tumour disappears altogether (complete response CR) then this is defined as an overall response (PR or CR). The treatment groups are compared with respect to overall response but in addition the duration of response is also investigated. In this respect, tumour progression is defined as a set increase in tumour volume or tumour re-appearance for a PR and CR respectively. Now, duration of response is often reported and analysed (Marty, 1997) and indeed this approach is supported by the WHO handbook for reporting results of cancer treatment (1979). This handbook states, that *when advanced breast cancer is being treated, it is necessary to determine the proportion of responders, the duration of responses, as well as survival.* However the analysis of duration of response is biased in terms of the randomisation since non-responders are excluded. An alternative valid analysis is to include all randomised subjects and compare the treatments with regard to time to progression. Subjects who do not respond will eventually progress as their tumour volume increases, and so are easily incorporated in the analysis of all randomised subjects.

2.6 DISCUSSION

It has been shown in this chapter how randomisation provides the basis for producing unbiased estimates of treatment differences from clinical trials, and how in order to retain this important attribute, data exclusion to form sub-populations or subgroups must be judged to be independent of the treatment assignment. Indeed examples have been

provided from a wide variety of therapeutic areas where this criterion has not been met and where the resulting estimates should be considered unreliable as a result.

From regulatory perspective, many of the principles expounded are directed towards the minimisation of bias at both the design and analysis stage. However while the stated aim is often to limit or minimise bias (ICH E9), there are specific areas at the analysis and reporting stage where perhaps the aim would more accurately be described as control of the direction of the bias such that estimated treatment differences are conservative for their intended purpose. In this context sensitivity analyses play an important and integral regulatory role since the assumptions behind the data conventions are typically unverifiable. Indeed this is why the *Points to Consider on switching between superiority and non-inferiority* (2000) is so compelling, as it brings to a head the different recommended approaches to be adopted for studies designed to show superiority compared with those designed to show equivalence or non-inferiority. In this respect has been shown that the accepted view that the ITT principle necessarily leads to a diminished estimate of the treatment difference is based on a series of strong assumptions regarding non-differential misclassification.

Despite concerted efforts to replace it, the ITT principle still proves popular and remains a convenient term to use, while the term Full Set has yet to fully establish itself despite the overall impact of ICH E9 on statistical practice. Indeed Simon Day (2002) - Lewis's replacement at the MHRA - has gone so far as to propose an end to the term ITT, which he regards - along with Per Protocol - as culpable *for ill-thought-out sensitivity analyses of a less than ideal study design.* Day's chosen emphasis now relates to *the importance of pragmatic trials rather than the importance of ITT analyses or populations.* However given the far reaching implications of the Schwartz and Lellouch philosophy regarding

pragmatic studies, including the absence of formal statistical analyses for the studies, it is unclear how this would fit into the requirements of drug regulation.

With regard to subject subgroups, regulatory guidance has been reviewed from a number of separate guidelines. In particular it has been shown how regulatory concerns regarding multiplicity can be addressed though the through pre-specification of subgroups together with more formal statistical approaches – such as the step-down procedure. It has also been highlighted how quantitative differences between subgroups with regard to the treatment effect may simply reflect the scale of measurement employed and that by implication qualitative interactions are of greater importance – particularly with regard to treatment policies. It has also been shown how subgroup inconsistencies are to be expected in the presence of modest overall treatment differences. As a consequence the investigation of differential treatment effects in subgroups is most rationally conducted through interaction analyses. In particular pre-specification of both the factor and the expected differential treatment effects in the resulting subgroups is important to reduce the opportunity for spurious findings and conclusions. In the next chapter (Chapter Three), the theme of observing inconsistent results when conducting subgroup analyses will be continued with the specific investigation of Simpson's paradox in randomised clinical trials - a reversal effect whereby the differences between treatments from all subgroups are in the opposite direction to the overall difference between treatments. In the following chapter (Chapter Four), the uniformity of treatment effect across subgroups will be considered through the investigation of treatment by subgroup interactions.

CHAPTER THREE: SIMPSON'S PARADOX AND RELATED INCONSISTENCIES

Old Ebeneezer
Thought he was Julius Caesar
And so they put him in a Home
Where they gave him medicinal compound
And now he's Emperor of Rome.

3.1 INTRODUCTION

Although rather rare, cases have appeared in the literature where the results of all subgroup analyses have indicated differences between treatments that are in the opposite direction to the overall difference between treatments (Julious and Mullee, 1994). This observation is known as Simpson's paradox (SP) after EH Simpson who discussed the problem in his paper *The interpretation of interaction in contingency tables* in 1951 (Simpson, 1951). However, according to Aldrich (1995), Karl Pearson and colleagues identified the phenomenon over fifty years earlier in their consideration of spurious correlations (Pearson K *et al*, 1898). As Aldrich points out, rather than considering *reversals of sign in dependence* their focus was directed towards *the mistaking of independence for dependence*. Since 1951, numerous publications have appeared on the topic of SP - most notably Blyth (1972), Lindley and Novick (1981), Zidek (1994), Aldrich (1995) and Pearl (2000). However reported cases of SP relate exclusively to observational studies and experiments that have employed randomisation appear untouched by the phenomenon. This chapter examines the potential for observing SP in the clinical trial setting and then uses the framework developed as a vehicle to examine less extreme – but related - inconsistencies that are more likely in randomised trials. As such, this chapter is about balance - the impact of baseline imbalance between treatment

groups when estimating treatment differences and the definition of balance itself. In contrast, this chapter is not about interactions; indeed the frameworks chosen deliberately exclude them. This is not to say that interactions are unimportant – although they are sometimes a simple artefact of the chosen scale of measurement – rather that their inclusion here would detract from the main theme. Instead treatment by subgroup interactions are addressed separately in the next chapter (Chapter Four).

In Section 3.2 of this chapter, the features of SP are described by way of an example. Simple probability theory is then used to provide a more formal framework to study the paradox. The potential for observing SP in the clinical trial setting is discussed in Section 3.3 and in this context the impact of randomisation and stratification are examined. In Section 3.4, a general mechanism for observing inconsistent results is presented. In Section 3.5, the more general case where the treatment effects in the subgroups are either all greater than or all less than the overall treatment effect is considered. The odds ratio model is given special focus and the impact of underestimation in the unconditional model is considered as is a re-definition of the concept of balance that restores the additive nature of the model. Finally, the results of some simulation exercises – aimed at establishing the chances of observing SP and related inconsistencies when randomisation is employed - are reported in Section 3.6 to support the earlier findings.

3.2 EXPLAINING SIMPSON'S PARADOX

Simpson's paradox is best described by considering an example from the literature. Table 3.I presents a historical comparison of two methods of kidney stone removal reported by Charig *et al* (1986). Overall, the percutaneous nephrolithotomy method (PN) of removal has a higher proportion of successes compared with the open surgery method (OS). However a comparison of the methods within each of two stone diameter subgroups

shows what appears to be a paradoxical result – that is, the OS method is more successful. Two features of the data are worth noting. Firstly, regardless of method of removal, a higher proportion of subjects with a stone diameter <2 cm is classified as a success and secondly the PN method was applied to a higher proportion of the subjects with a stone diameter <2 cm compared with the OS method. The explanation of this second feature in this historical comparison is that the choice of method is influenced by the stone diameter of the subject. For a randomised study, however, where subjects would be randomised to the method of removal, this explanation would be implausible.

Table 3.I. An example of Simpson's paradox

Method of removal	Stone diameter		Total
	<2 cm	>=2 cm	
Open surgery, 1972-80	81/87 = 0.93	192/263 = 0.73	273/350 = 0.78
Percutaneous nephrolithotomy, 1980-85	234/270 =0.87	55/80 = 0.69	289/350 = 0.83

Simpson's paradox can be illustrated using simple probability theory as described by Hand (1994). Let there be two treatments (T=1,2), two possible treatment outcomes (X=1,2) and two possible levels of a factor (F=1,2) such that two mutually exclusive subgroups can be defined.

Using basic probability axioms it is simple to show that for T=1:

$$p(X=1|T=1) = p(X=1|F=1, T=1)\, p(F=1|T=1) + p(X=1|F=2, T=1)\, p(F=2|T=1) \qquad (1)$$

Similarly for T=2:

$$p(X=1|T=2) = p(X=1|F=1, T=2)\, p(F=1|T=2) + p(X=1|F=2, T=2)\, p(F=2|T=2) \qquad (2)$$

Now, SP describes the situation where $p(X=1|T=1) < p(X=1|T=2)$ but both $p(X=1|F=1, T=1) > p(X=1|F=1, T=2)$ and $p(X=1|F=2, T=1) > p(X=1|F=2, T=2)$. (Or alternatively, where $p(X=1|T=1) > p(X=1|T=2)$ but both $p(X=1|F=1, T=1) < p(X=1|F=1, T=2)$ and $p(X=1|F=2, T=1) < p(X=1|F=2, T=2)$).

The apparent paradox can easily be explained by considering the weighting system employed. Equations (1) and (2) can be re-written with weights ω_1 and ω_2 corresponding to the probabilities $p(F=1|T=1)$ and $p(F=1|T=2)$ respectively.

$$p(X=1|T=1) = p(X=1|F=1, T=1)\, \omega_1 + p(X=1|F=2, T=1)\, (1 - \omega_1) \qquad (3)$$

$$p(X=1|T=2) = p(X=1|F=1, T=2)\, \omega_2 + p(X=1|F=2, T=2)\, (1 - \omega_2) \qquad (4)$$

If $\omega_2 >> \omega_1$ and both $p(X=1|F=1, T=1) >> p(X=1|F=2, T=1)$ and $p(X=1|F=1, T=2) >> p(X=1|F=2, T=2)$ then the term $p(X=1|F=1, T=2)\, \omega_2$ will dominate and lead to a reversal of the overall treatment difference compared with the treatment differences within the two subgroups. This can be shown numerically by substituting the data from Table 3.I in equations (1) and (2) to give:

$$p(\text{success} \mid \text{OS method}) = 0.93 \times 0.25 + 0.73 \times 0.75 = 0.78$$

$$p(\text{success} \mid \text{PN method}) = 0.87 \times 0.77 + 0.69 \times 0.23 = 0.83$$

The paradox can easily be avoided by using identical weights ($\omega_1 = \omega_2 = \omega$) when calculating each treatment effect. This produces an overall treatment difference that is in

the same direction as the within subgroup treatment differences. That is, the difference p(X=1|T=1) - p(X=1|T=2) which is calculated as:

$$[p(X=1|F=1, T=1) - p(X=1|F=1, T=2)]\omega + [p(X=1|F=2, T=1) - p(X=1|F=2, T=2)](1-\omega)$$

must be negative if

$$p(X=1|F=1, T=1) < p(X=1|F=1, T=2) \text{ and } p(X=1|F=2, T=1) < p(X=1|F=2, T=2)$$

The simplest weighting system to employ is to use $\omega = (1 - \omega) = ½$ that weights each subgroup equally and consequently takes no account of the number of subjects in each subgroup. Using the data from Table 3.I again and employing the weights $\omega = (1 - \omega) = ½$ shows how the paradox can be avoided.

p(success | OS method) = 0.93 x 0.5 + 0.73 x 0.5 = 0.83

p(success | PN method) = 0.87 x 0.5 + 0.69 x 0.5 = 0.78

Perhaps the most common method is to use weights such that the variance of the contrast is minimised. That is, $\omega \propto n_{11} n_{21}/(n_{11} + n_{21})$ and $(1 - \omega) \propto n_{12} n_{22}/(n_{12} + n_{22})$ where n_{ij} represents the number of subjects in each combination of treatment (i=1,2) and factor (j=1,2). The choice of which weights to use, to combine the parameter estimates from the separate subgroups, relates directly to the question posed and is not arbitrary (Hand, 1994; Lane and Nelder, 1982). Arguably the first consideration is whether the question posed is a conditional one, and if it is, then the second consideration is the exact nature of the conditional question. Thus if interest lies in the unconditional effect of treatment then the weights ω_1 and ω_2 are simply the probabilities p(F=1|T=1) and p(F=1|T=2) as described in

the earlier example. In this respect each subject is weighted equally and the factor is ignored. Alternatively if a conditional question is posed then the weights applied will need to be directed towards the specific formulation of the problem. For instance, if one were interested in the effect of treatment in a defined population then weights would be selected according to the know proportions in the population with regard to the factor of interest. Other alternatives could include using the observed proportions in the data, or equal weights for each level of the factor (for instance to predict the effect of treatment had all centres randomised the same number of subjects). Such approaches are common in epidemiology and demography where the term standardisation is used to describe factor adjustment (Lane and Nelder, 1982). The choice of weights is discussed further in Section 3.8 in the context of prediction.

3.3 THE IMPACT OF RANDOMISATION AND STRATIFICATION

As described in Chapter Two, in the clinical trial setting, one aim of randomisation is to balance treatment groups for treatment independent variables or factors – known and unknown, measured and not measured – which may influence outcome (Gillings and Koch, 1991). It follows that randomisation allows valid inference over all possible random assignments and this is the principle that underpins the randomised and controlled trial (RCT). Furthermore, subgroups formed on the basis of treatment-independent factors, generate treatment groups that will – over all randomisations – still be balanced and in this context treatment comparisons within these subgroups are unbiased. Tests can be performed within subgroups as though randomisation had been performed separately within those subgroups (Lachin, 1998), the resulting test statistics are independent and as such can be combined across levels of a factor as if stratification had been performed. That is not to say, however, that the formation of subgroups has no impact since it modifies the populations to which the results apply and in these different subgroups the

effects of the treatments may indeed differ. To examine the influence of randomisation in generating consistent results when overall and subgroup analyses are compared it is useful to adopt the probability framework described in the preceding section.

Consider a RCT with unrestricted randomisation where r is the odds of assignment to treatment 1 compared to treatment 2. That is, $p(T=1) = r\,p(T=2)$. Factor F is recorded for all subjects prior to randomisation and can take one of two values. As assignment to treatment is independent of factor F then $p(T=1) = p(T=1|F=1)$ and $p(T=2) = p(T=2|F=1)$ which leads to $p(T=1|F=1) = r\,p(T=2|F=1)$. It is then simple to show using Bayes' theorem that $p(F=1|T=1) = p(F=1|T=2)$. That is, the paired weights ω_1 and ω_2 from (3) and (4) are equal producing a theoretical basis for randomisation providing protection against SP.

Prospective stratification by F in an RCT effectively forces balance with regard to F and limits the scope for variability around the expectation, since subjects are randomised separately for each level of F using fixed or variable length blocks. That is, $p(T=1|F=1) = r\,p(T=2|F=1)$ is well controlled with the result that $p(F=1|T=1) = p(F=1|T=2)$ is also controlled. Consequently stratification effectively eliminates the chance of observing SP when subgroup analyses are generated from prospectively stratified factors.

These findings can be contrasted with those expected from other types of medical research such as epidemiology where the independence of treatment and factor can not be guaranteed in the typical situation where randomisation has not been employed. In these cases, confounding can be a real concern and serious bias can result. It is not surprising therefore that reported occurrences of SP have been reserved for non randomised, observational studies.

Pearl (2000) considers SP in relation to causality and introduces a distinction between *seeing* and *doing*. In this respect, Pearl's *do-operator* represents the causal condition, *given that we do*, which can be contrasted with the more usual condition, *given that we see*. It follows that the inequality, $P(E \mid C) > P(E \mid \neg C)$, where C refers to cause (test treatment, say), E to effect, and ¬C to the complement of C (reference treatment say), is not actually stating that C is a positive causal factor of E. Rather, although C is positive evidence for E, confounding factors may in fact be causing both C and E. Instead, to represent C as the positive causal factor it is necessary (according to Pearl's terminology) to write the inequality as $P(E \mid do(C)) > P(E \mid do(\neg C))$. Now, it is clear that in a properly conducted RCT, such potential confounding factors cannot influence the assignment of randomised treatment and it is appropriate to use the do-operator. Indeed in relation to causality, this is why the RCT is so powerful since it is the long run balancing of treatment groups with regard to factors through randomisation that enables cause and effect to be assigned to treatment with confidence. (Note that Pearl actually regards randomisation as a causal concept as opposed to a statistical one, such as likelihood or conditional independence.) That is not to say that the do-operator is automatically applicable to subgroups in a RCT since randomisation can not balance groups for variables or factors that are not independent of treatment – for example, subgroups constructed on the basis of drug compliance. (This point - that subject exclusion on the basis of treatment dependent data can introduce bias - was discussed extensively in Chapter Two.) Furthermore, Pearl highlights the problem of intermediate events that also reside on the causal pathway in clinical trials. An example of this would be the sub-grouping of a long-term primary outcome measure in HIV/AIDS, such as survival, by a short-term surrogate marker of the disease, such as a post-treatment level of the CD4 count. In this case, since randomised treatment affects the post-treatment CD4 count, randomisation provides no protection

against SP and the unconditional estimate of the treatment difference is the most appropriate.

3.4 A MECHANISM FOR OBSERVING INCONSISTENT RESULTS

Although over all randomisations treatment groups will be balanced in terms of the distributions of pre-randomisation or baseline factors, observed randomisations will exhibit at least some imbalance and as a consequence it is important to examine the potential impact of such imbalance on estimates of the treatment difference. The following section describes a mechanism for observing inconsistent results between overall and subgroup analyses in the presence of observed imbalance when the outcome variable (y) is continuous. To illustrate the mechanism a linear model is used which includes terms for factor and treatment but assumes no treatment by factor interaction.

As in the previous illustrations, consider a RCT with two treatment groups, T=1 and T=2 and a factor F that can be used *a posteriori* to form two subgroups, F=1 and F=2. It is assumed that a subject (h) is subject to the effects of treatment τ_i (i=1,2) and factor ϕ_j (j=1,2) and that there is also a background level, represented by the constant term α. This leads to a model of the form:

$$y_{ijh} = \alpha + \tau_i + \phi_j + \varepsilon_{ijh}$$

where the errors ε_{ijh} are independent and identically distributed with constant variance and where to estimate the parameters, constraints must be applied to their estimates (Nelder, 1994a). In this respect the parameters τ_i and ϕ_j can be regarded as expected values for the effects of treatment and factor respectively and an individual can also be thought of as having an expectation equal to $\alpha + \tau_i + \phi_j$.

The impact of factor imbalance on the estimates of the treatment difference can be examined by working with the expected value for each subject. Table 3.II gives the expectation $(\alpha + \tau_i + \phi_j)$ for a subject classified to each one of the four treatment by factor combinations together with the resulting number of subjects (n_{ij}) in each combination following unrestricted randomisation to treatment.

Table 3.II. Expected outcome values for two treatment, two factor design

	Factor F=1	Factor F=2
Treatment T=1	$\alpha + \tau_1 + \phi_1$ (n_{11})	$\alpha + \tau_1 + \phi_2$ (n_{12})
Treatment T=2	$\alpha + \tau_2 + \phi_1$ (n_{21})	$\alpha + \tau_2 + \phi_2$ (n_{22})

To estimate the treatment difference in each subgroup the mean value for T=2 in the given subgroup is subtracted from the corresponding mean value for T=1 which gives $(\tau_1 - \tau_2)$.

In contrast, an overall estimate of the treatment difference which ignores factor F is simply calculated by subtracting the mean score across all $(n_{21} + n_{22})$ observations for T=2 from the mean score across all $(n_{11} + n_{12})$ observations for T=1. This is given as follows:

$$(\tau_1 - \tau_2) + [n_{11}/(n_{11} + n_{12}) - n_{21}/(n_{21} + n_{22})](\phi_1 - \phi_2)$$

$$= (\tau_1 - \tau_2) + \kappa(\phi_1 - \phi_2) \text{ where } |\kappa| \leq 1 \qquad (5)$$

What this simple illustration shows is that in the presence of factor imbalance the unconditional estimate of the treatment difference will be contaminated by a proportion of the difference between the effects of F=1 and F=2. The term "contamination" has been coined to distinguish this effect from the term bias. Bias is a systematic effect on a parameter estimate whilst in this context contamination can be thought of as a non

systematic effect which would not lead to bias since over all randomisations the expectation of the coefficient κ would equal 0. This reinforces the point that the overall (unconditional) estimate of the treatment difference is unbiased and that imbalance manifests itself in terms of the variability of the estimate (Senn, 1994).

For an observed randomisation, contamination will be zero in this framework if one or both of the following conditions hold:

- Factor F has no influence on outcome. That is, $(\phi_1 - \phi_2) = 0$.
- In each treatment group, the observed proportion of subjects with characteristic F=1 is identical such that $\kappa = 0$. That is, $n_{11}/(n_{11} + n_{12}) = n_{21}/(n_{21} + n_{22})$.

These results correspond to those obtained earlier but enable the impact of imbalance to be quantified in a meaningful way for continuous outcomes. To assess the actual impact of imbalance it is worth constructing a simplified example where (z)% of subjects in T=1 and $(100 - z)$% of subjects in T=2 have characteristic F=1 and where the number of subjects randomised to each treatment group is equal. This leads to the simplified conditions, $n_{11} = n_{22}$ and $n_{12} = n_{21}$ in which case κ in (5) reduces to $(n_{11} - n_{12})/(n_{11} + n_{12})$. This allows one to work with one measure of imbalance, $|2z - 100|$%, in both treatment groups such that $\kappa = |2z - 100|/100$. Some examples of the values of κ that are produced for increasing degrees of imbalance are given below. For instance a modest imbalance of 10% leads to $\kappa = 0.1$ which would produce contamination equal to 10% of the differential effect of factor on outcome.

Imbalance (z)% versus $(1-z)$%	Proportion (κ)
50% versus 50%	0
55% versus 45%	0.1
65% versus 35%	0.3
80% versus 20%	0.6

In this framework, for SP to exist it is a necessary condition that the differential effect of factor F is greater than the differential effect of treatment T. That is, $|\phi_1 - \phi_2| > |\tau_1 - \tau_2|$, although note in this respect that variability in the realisations of τ and ϕ is deliberately ignored here. Also, unless the factor imbalance is dramatic then the differential effect of the factor must be substantially greater than the differential effect of treatment. As such, randomised studies that employ an active control or reference group and are designed to show therapeutic equivalence (such that $|\tau_1 - \tau_2| < \delta$, where δ is considered to be clinically unimportant) would, *a priori*, seem to be the most vulnerable to SP.

Consistency between the individual subgroup differences and the overall treatment difference can be achieved by adopting a suitable set of weights to combine the estimates from each subgroup in a similar manner to that described in section 3.2. The simplest approach is to weight the differences between treatments from the subgroups equally such that $\omega = (1 - \omega) = ½$. An alternative system is to use weights that minimise the variance of the contrast such that $\omega \propto n_{11} n_{21}/(n_{11} + n_{21})$ and $(1 - \omega) \propto n_{12} n_{22}/(n_{12} + n_{22})$. This system gives the most weight to the treatment difference for the subgroup that contains the greatest number of subjects and within each subgroup the weight is maximised when the treatment groups have an equal number of subjects assigned. It is interesting to note that for the earlier illustrations where (z)% of subjects in T=1 and (100 -z)% of subjects in T=2 had characteristic F=1, and where $n_{11} = n_{22}$ and $n_{12} = n_{21}$, then the two weighting systems would lead to equivalent results since the total number of subjects and the degree of imbalance in each subgroup is identical in each case. (See Table 3.III later for an example.)

From a computing perspective, the system of equal weights corresponds to the system employed by the SAS procedure GLM (SAS, 1989) when, in addition to the main effects

of factor and treatment in the model, an interaction term between treatment and factor is also included in the model statement and type III sum of squares (SS) is requested. The weighting system that minimises the variance of the contrast is also available in SAS and can be obtained in a number of ways. If the model includes just factor (specified first in the model statement) and treatment then the type I SS option will produce the required output. Alternatively types II, III or IV SS produce an identical result with the same model statement but in this case the ordering of the two terms is unimportant. Nelder (1994a) provides a thorough review of the SAS procedure GLM and in particular a highly critical review of type III SS option.

In the context of the binary outcome model, Nelder (1994b) has indicated that SP always requires an interaction term in the model since the margins do not provide an adequate summary of the content of the table. The context of his comment is not entirely clear and Nelder may have simply been referring to a log linear model formulation. Nevertheless, what this formulation shows is that for continuous outcomes, this is not necessarily the case and that SP (and other less extreme inconsistencies) could be explained adequately with a model excluding an interaction term and consequently one that satisfies the parsimony test. This is also true for binary data when the logistic regression model is applied.

This formulation also shows that it is just as likely that imbalance would lead to inflation of the estimate of the treatment difference as it is that it would lead to a reversal of the treatment effect as in SP. This would occur if the observed imbalance were in the opposite direction. Indeed the mechanism described above can be used to assess the effects of factor imbalance in a general manner and this is discussed in the following section.

3.5 LESS EXTREME INCONSISTENCIES

It is a commonly held view – see for instance Peto (1982), Koch (1996) and Senn (1997) - that if the treatment difference were observed to be larger in one subgroup of subjects compared to the overall treatment difference then the treatment difference in the complement subgroup would be smaller. Indeed, Koch describes this as one reason for not adjusting the p-values from subgroup analyses for multiplicity. The reasoning is that the regulatory authorities who assess drug applications look for homogeneity of effect and subgroup analyses are likely therefore to lead to a weakening of the treatment conclusion rather than a strengthening of it. (For instance, if the overall treatment effect is statistically significant but inconsistent results are observed across subgroups in terms of statistical significance, support for the conclusion of homogeneity of effect is reduced.) Whilst it is typically observed that the overall result lies in between the subgroup results, the assertion assumes that there is no notable imbalance with regard to treatment assignment in each subgroup and as a consequence the point lends itself to examination using model (5). First consider a fictitious example where this does not happen.

Consider a RCT where outcome is diastolic blood pressure (DBP) recorded following treatment with either active or placebo treatment. The results of trial are given in Table 3.III where mean DBP [mmHg] is summarised by treatment and also by two age subgroups. The number of subjects in each cell is given in parenthesis. The difference between the treatment means (placebo - active) is 10 mmHg in both subgroups whereas the overall treatment difference is 8 mmHg, 2 less than the subgroup differences. There is no evidence of an interaction between treatment and the age classification but there is an imbalance in the proportion of subjects aged <65 years in the treatment groups.

Table 3.III. An example where the treatment effect in both subgroups is larger than overall treatment effect

	<65 years	≥65 years	Total
Active	90 (N=40)	100 (N=60)	96 (N=100)
Placebo	100 (N=60)	110 (N=40)	104 (N=100)

Using model (5) it is simple to see that if the form of the model is appropriate - in that it is considered to provide an adequate description of the data - then the apparent inconsistency, as given in Table 3.III, should not be unexpected. That is, in the absence of an interaction but in the presence of an imbalance ($0< \kappa< 1$), if treatment and factor have independent effects on outcome ($|\tau_1 - \tau_2| > 0$ and $|\phi_1 - \phi_2| > 0$) then the expected treatment difference for both subgroups would differ from the overall result to the same extent and in the same direction. Of course this formulation ignores the inherent variability associated with the outcome variable which leads to a reduction in the chances of this phenomenon occurring in practice.

3.6 RANDOMISATION AND THE ODDS MODEL

Now, consider the odds model for binary outcome data. It has been shown that even a balanced design, where a factor is perfectly balanced across treatment groups, does not lead to identical conditional and unconditional estimates of the treatment difference. That is, balance actually leads to underestimation of the unconditional treatment effect if the factor excluded from the analysis has an independent impact outcome. In this case the odds ratio shrinks towards unity.

Gail (1986) gives a general formula for the approximate asymptotic bias of the treatment effect for the exponential family of models including the logistic model. As an illustration consider Table 3.IV where for each treatment by factor combination the proportion of

successes is presented together with - in parenthesis - the associated odds. The trial is perfectly balanced in the traditional sense in that there is an identical number of subjects (N=100) for each combination of treatment and factor.

Table 3.IV. An example of underestimation of the unconditional odds ratio when a trial has perfect balance

	Factor F=1	Factor F=2	Total
Treatment T=1	90/100 *(9)*	75/100 *(3)*	165/200 *(4.71)*
Treatment T=2	75/100 *(3)*	50/100 *(1)*	125/200 *(1.67)*
Odds ratio	3	3	2.83

Within each subgroup the odds ratio (T=1/T=2) is 3 indicating a constant treatment difference with no treatment by factor interaction, but despite the apparent balance of the trial the overall odds ratio of 2.83 is smaller. (Note that the same table was used in Chapter Two (Table 2.I) to illustrate how an apparent treatment by factor interaction can disappear through transformation of the scale - in this case from proportions to odds.)

It is straightforward to derive a simple formula for calculating the unconditional odds ratio (ψ) from a combination of the four separate odds (λ_{ij}) if it is assumed that the number of subjects in each treatment (i=1,2) by factor (j=1,2) combination is identical. (Refer to Appendix A for the derivation.) This gives,

$$\psi = (\lambda_{11} + \lambda_{12} + 2\lambda_{11}\lambda_{12})(2 + \lambda_{21} + \lambda_{22}) / (\lambda_{21} + \lambda_{22} + 2\lambda_{21}\lambda_{22})(2 + \lambda_{11} + \lambda_{12}) \qquad (6)$$

that produces a value of 2.83 when the four odds from Table 3.IV are substituted into (6). For a given treatment effect, underestimation increases as the size of the factor effect increases. Furthermore, the extent of the underestimation diminishes as the treatment odds

ratio approaches either the relative risk (π_1/π_2) or the inverted reverse relative risk ($1-\pi_2$)/($1-\pi_1$) and due to the symmetric nature of the odds ratio, this occurs as the odds approach either zero or infinity. (Further detail relating to underestimation of the odds ratio is provided in Chapter Five, Section 5.5.) However since the relative risk (ratio of proportions) behaves in a similar way to model (5) for continuous data, as the odds ratio approaches the relative risk (or inverted reverse relative risk) the additive nature of the unconditional analysis returns. Indeed, in epidemiological research the odds ratio is often regarded as an approximate relative risk since incidence and prevalence rates are frequently low and denominators are large. However in drug development this is usually not the case.

That is not to say that the randomisation principle does not hold for the odds model since the true treatment difference is simply reduced rather than eliminated. Furthermore, under the null hypothesis of no treatment effect, the expected value of the odds ratio remains 1 regardless of whether a factor effect exists or not. Of course, in practice, observed imbalance can impact overall analyses in much the same way as previous models – that is, it can lead to both under and over estimation of the true treatment difference.

Interestingly, it is possible to achieve consistent results with the odds model but this requires a re-definition of the concept of balance. Consider the case where balance is re-defined in terms of the proportion of successes with characteristic F=1 in each treatment group rather than proportion of subjects *per se*. (Note that due to symmetry considerations this condition could also be defined in terms of the number of failures.) Table 3.V provides an example where this balance redefinition criterion is met. For treatment T=1, the number of successes with characteristic F=1 as a proportion of the total number of successes for treatment T=1 is 0.455 (50/110). Similarly for T=2 the corresponding

proportion is 0.455 (75/165). If failures are considered instead, then the proportions are again identical – that is, 0.667 (400/600) and 0.667 (200/300) for T=1 and T=2 respectively. As the data in Table 3.V demonstrate, the odds ratio in each subgroup is identical to the overall odds ratio despite factor F having an independent effect on outcome. The proof of this result is shown in Appendix B.

Table 3.V. An example of consistency between the unconditional odds ratio and subgroup odds ratios when balance is re-defined.

	Factor F=1	Factor F=2	Total
Treatment T=1	50 / 450 *(0.125)*	60 / 260 *(0.3)*	110 / 710 *(0.183)*
Treatment T=2	75 / 275 *(0.375)*	90 / 190 *(0.9)*	165 / 465 *(0.550)*
Odds ratio	3	3	3

Unfortunately this re-definition is not helpful in the practical sense since no mechanism exists which could provide this balance for all factors which could impact on outcome since it involves balancing the outcome which is unpredictable rather than the baseline variables which are known at the time of balancing.

At this stage, and for completeness, it is useful to consider some weighting systems that are employed to combine the individual estimates of the odds from subgroups to form an estimate of the common treatment odds ratio that avoids SP. The first example actually combines odds ratios from the subgroups that have been transformed using natural logarithms to form an estimate of the common odds ratio. For the case were there are two treatments (i=1,2) and a factor with two levels (j=1,2) the common odds ratio (Ψ) is estimated as follows:

$$\log(\psi) = \frac{\omega_1 L_1 + \omega_2 L_2}{\omega_1 + \omega_2}$$

where $L_j = \log\left[\frac{p_{1j}(1-p_{2j})}{p_{2j}(1-p_{1j})}\right]$ and $\frac{1}{\omega_j} = \frac{1}{n_{1j}p_{1j}} + \frac{1}{n_{1j}(1-p_{1j})} + \frac{1}{n_{2j}p_{2j}} + \frac{1}{n_{2j}(1-p_{2j})}$

In this case the reciprocal of the weight $(1/\omega_j)$ is the squared standard error of L_j and it follows that the larger the standard error, the smaller the weight.

A popular alternative estimate of the common odd ratio, proposed by Mantel and Haenszel (1959), does not in fact actually weight the individual odds ratios. Rather identical weights are assigned to each set of terms $p_{1j}(1-p_{2j})$ and $p_{2j}(1-p_{1j})$. These terms are the summed across all levels of the factor prior to forming the required ratio. This is illustrated below,

$$\psi = \frac{\omega_1'[p_{11}(1-p_{21})] + \omega_2'[p_{12}(1-p_{22})]}{\omega_1'[p_{21}(1-p_{11})] + \omega_2'[p_{22}(1-p_{12})]}$$

where $\omega_j' = \frac{n_{1j}n_{2j}}{n_{1j}+n_{2j}}$

It is straightforward to imagine examples where some observed proportions are either zero or one, in which case the Mantel-Haenszel and back-transformed log odds ratios will obviously differ since the former is able to utilise more information. As such the Mantel-Haenszel approach has been recommended for cases were the number of subgroups (or strata) is large but the number of observations in each subgroup is relatively modest (Fleiss, 1986). (For instance when investigator site is the strata and few subjects are randomised at each site.) In contrast, the logarithmic approach is recommended in cases were the number of subgroups is small and number of observations within each subgroup is large – in this instance it is deemed to perform better, or at worst, only a little poorer than the alternatives. It also has the advantage that it is easily extended to multi-treatment

and multi-factor problems through the logistic model. (Refer to Chapter Five, sections 5.3 and 5.5 for further discussion of the logistic model.)

Fleiss (1986) provides further details of this and other methods of estimating a common odds ratio and also gives comprehensive details of the weights that can be used for differences in means and differences in proportions.

3.7 SUPPORTING SIMULATIONS

To examine the behaviour of the overall difference between two treatments relative to that of the subgroup treatment differences in a RCT setting a series of simulations were performed. A Normally distributed outcome was considered initially. This was followed by the more complex situation regarding a binary outcome.

Each simulation assigned subjects at random to one of two levels of a factor F using a binomial distribution. However since in practice most randomisation schema are blocked to ensure an approximate equal number of subjects to each treatment group, assignment of subjects to treatment was on a simple alternating basis. This simulated a parallel group study with two treatment groups and blocked randomisation. Each simulation was conducted on 5,000 trials using a random number generator from the uniform distribution in SAS to assign the subject to one of two levels of the factor F using the 0.5 cut-off. The resulting incidence percentages have been reported to two decimal places since it was important to identify any occurrence of SP under the conditions tested. In this respect, a reduction in the number of decimal places would have meant that some combinations would have been reported as having an incidence of zero even though SP had been observed. To aid interpretation, standard errors (SE) calculated using the Normal approximation to the binomial distribution are given below for a range of values - since the

magnitude of the Monte Carlo error varies according to the value of the parameter being estimated. Note that in cases where a zero incidence was observed, the exact 95% confidence interval (StatXact software: Cytel, 1999) was 0 to 0.07%. Further details are provided in the Simulation Note at the end of this Research Thesis.

Incidence %	0.02	0.1	0.2	0.5	1.0	2.5	5.0	7.5	10	20	30	40	50
SE %	0.020	0.045	0.063	0.10	0.14	0.22	0.31	0.37	0.42	0.57	0.65	0.69	0.71

3.7.1 Normally distributed outcome

For the normally distributed case, subject outcome was determined through the assignment of a Normal distribution to each of the four treatment by factor combinations. The parameters of Normal distribution were chosen such that both treatment and factor had an independent effect on outcome. That is, the treatment difference for F=1 was identical to the treatment difference for F=2. With regard to standardised treatment differences - that is, $\frac{|\tau_1 - \tau_2|}{\sigma}$, Cohen (1997) suggests a range of 0.1 to 1.0 where 0.2 might represent a small effect and 0.8 a large effect. Therefore, in this series of simulations the effects ranging from zero to 2.0 (0, 0.1, 0.2, 0.3, 0.4, 0.5, 1.0 and 2.0) have been selected in order to provide a broad coverage of plausible differences. Identical treatment and factors effects have been chosen to provide symmetric comparisons in each case. The first six effects range from zero to 0.5 in increments of 0.1 to cover a range of small to medium size effects, including no effect. The last two effects of 1.0 and 2.0 represent large and very large effects respectively to investigate the impact at the limit of plausibility. For instance, Machin and Campbell's (1987) sample size tables for the difference between two means terminates at 1.5. A reference mean was selected which fixed the mean for the T=1, F=1 combination of treatment and factor at 0 while the standard deviation was fixed at 1 for all 4 combinations. For example, for a treatment effect of 0.3 and a factor effect of 0.2 the following Normal distributions (N_{TF}) were selected: $N_{11}(0,1)$, $N_{12}(0.2,1)$, $N_{21}(0.3,1)$,

$N_{22}(0.5,1)$. The effect of sample size was also investigated using three different scenarios corresponding to small, medium and large sized clinical trials. That is, 40, 200 and 1000 subjects. Forty subjects is most likely to represent a small phase II study while 200 subjects could represent a large phase II or small phase III study. The large study with 1000 subjects (500 per treatment group) represents a substantial phase III study - although some phase III studies are certainly larger than this. These three scenarios resulted in a total of 192 (8x8x3) simulations being performed.

As the relative sizes of the treatment and factors effects varied, two features were examined. First the potential for observing SP, and second the proportion of cases where the overall difference between the treatment means was either greater than or less than the treatment differences for both subgroups. The results of the simulation are given in Table 3.VI.

Table 3.VI. Simulation 1: Normally distributed outcomes: Percentage where overall treatment difference > or < treatment effect in both subgroups (% Simpson's paradox) with varying sample size

		Mean factor difference							
Mean treatment difference	Total sample size	0	.1	.2	.3	.4	.5	1.0	2.0
0	40	**7.84**	8.44	9.74	10.74	12.38	15.84	28.28	49.22
		(0.30)	(0.48)	(0.66)	(0.72)	(0.90)	(1.00)	(3.46)	(8.88)
	200	**3.88**	4.24	6.84	9.36	12.94	15.54	29.90	50.72
		(0.08)	(0.12)	(0.12)	(0.32)	(0.70)	(1.02)	(3.18)	(8.48)
	1000	**1.46**	2.76	6.46	9.82	12.88	15.98	29.90	49.50
			(0.08)	(0.20)	(0.38)	(0.64)	(1.06)	(2.86)	(7.94)
.1	40	7.22	**7.68**	9.64	11.20	13.58	15.68	28.48	48.76
		(0.56)	**(0.40)**	(0.38)	(1.06)	(0.84)	(1.16)	(3.78)	(8.18)
	200	3.88	**4.58**	6.52	9.16	11.92	15.82	30.00	48.54
		(0.06)	**(0.12)**	(0.26)	(0.38)	(0.46)	(0.74)	(2.72)	(7.68)
	1000	1.56	**3.32**	6.12	9.10	11.74	15.66	29.44	49.62
			(0.02)	(0.08)	(0.08)	(0.20)	(0.24)	(0.92)	(5.56)
.2	40	8.28	7.96	**9.52**	11.72	12.98	15.78	28.44	48.18
		(0.38)	(0.26)	**(0.38)**	(0.54)	(0.58)	(1.24)	(3.22)	(7.90)
	200	3.48	3.88	**6.64**	9.38	11.80	16.14	28.64	49.24
			(0.02)	**(0.18)**	(0.12)	(0.22)	(0.54)	(1.32)	(5.64)
	1000	1.80	3.26	**6.16**	9.38	12.56	15.12	29.04	49.58
						(0.04)	(0.04)	(0.08)	(0.94)
.3	40	7.78	8.02	9.70	**10.16**	12.62	15.00	28.80	48.40
		(0.32)	(0.30)	(0.44)	**(0.44)**	(0.66)	(0.82)	(2.58)	(7.04)
	200	3.48	4.24	6.44	**9.12**	13.44	15.34	29.06	48.74
		(0.02)		(0.06)	**(0.04)**	(0.04)	(0.12)	(0.92)	(3.44)
	1000	1.66	3.24	6.82	**9.78**	12.22	15.12	28.88	50.46
									(0.06)
.4	40	7.82	8.30	9.12	11.40	**13.08**	15.28	27.96	49.34
		(0.06)	(0.06)	(0.10)	(0.42)	**(0.58)**	(0.66)	(2.00)	(6.36)
	200	3.52	4.12	6.90	9.32	**12.40**	15.76	29.76	48.52
				(0.02)		**(0.02)**	(0.02)	(0.10)	(1.46)
	1000	1.34	3.32	6.22	9.28	**12.16**	15.18	29.32	50.56
.5	40	7.80	8.16	9.72	10.40	13.68	**15.94**	28.34	49.08
		(0.04)	(0.20)	(0.18)	(0.12)	(0.22)	**(0.46)**	(1.30)	(5.70)
	200	4.02	4.84	7.00	9.22	12.34	**15.12**	29.76	50.16
				(0.02)			**(0.02)**	(0.02)	(0.48)
	1000	1.40	3.18	6.28	9.28	12.02	**15.32**	29.10	48.80
1.0	40	8.46	8.48	9.14	10.60	12.36	15.96	**28.58**	48.46
		(0.02)					(0.02)	**(0.04)**	(0.72)
	200	3.68	4.68	6.58	9.40	13.50	15.68	**30.12**	49.10
	1000	1.54	3.12	6.40	8.64	11.72	16.04	**28.70**	48.68
2.0	40	7.72	8.60	9.10	10.62	13.40	15.42	28.30	**47.52**
	200	3.60	4.08	6.52	9.36	12.44	15.28	30.08	**50.34**
	1000	1.80	3.60	5.86	9.50	12.66	16.22	28.42	**47.96**

5000 simulated data sets with randomly assigned to F (p=0.5) but with balanced treatment groups

Simpson's paradox

As expected, the incidence of SP was low under the range of conditions selected with the maximum incidence (8.88%) occurring at the limits of the conditions investigated – that is, no treatment difference, the largest factor effect (2.0) and the smallest sample size (N=40). For a given treatment difference (table row), the incidence of SP increased with increasing size of factor effect – although when the treatment difference were high (2.0) and / or for larger sample sizes, SP was sometimes not observed for the range of factor effects investigated. For a given combination of treatment and factor effect, the incidence of SP decreased with increasing sample size, while for a fixed factor effect (table column) and fixed sample size, the incidence of SP decreased with increasing treatment difference. When the treatment and factors effects were identical (table diagonal bolded), the incidence of SP generally decreased with increasing standardised effects. That is, as $\frac{|\tau_1 - \tau_2|}{\sigma}$ and $\frac{|\phi_1 - \phi_2|}{\sigma}$ increased. However the trend was not entirely clear-cut with an indication that the incidence of SP may have increased initially before decreasing again – perhaps a result of the interplay between the increasing factor effect that increases the chances of SP and an increasing treatment effect that in contrast reduces them.

Under the condition of a non-zero treatment difference, the subgroups were most likely to show the correct direction of the treatment difference – rather than the unconditional treatment difference - when the factor effect was large and so was the sample size. For example, when the treatment difference was 0.1 and the factor difference was 2.0, then 251 of the 278 cases of SP observed showed the correct direction of the treatment difference in the subgroups when the sample size was 1000. However, as the both sample size and factor effect decreased, the assignment of correct direction became more balanced between the subgroups and the unconditional treatment difference.

Overall effect greater than or less than both subgroups

When the more general inconsistency, of the overall treatment effect being either greater than or less than both subgroups, was investigated it was found that as expected the size of treatment difference appeared to have no influence on observed incidence. However both the size of factor difference and the sample size did have an influence. As with SP, the incidence increased (table row) with increasing factor difference – indeed when the factor difference was at its maximum for the conditions investigated (2.0), the incidence reached was as high as 50%. When the factor effect was small, increasing the sample size reduced the incidence of this inconsistency. However for factor differences of 0.4 or greater, the incidence appeared largely unaffected by sample size for the range of values investigated.

In contract to SP, when the treatment and factors effects were identical (table diagonal), the incidence of the more general inconsistency clearly increased with increasing standardised effects. In this cases the increasing factor effect increasing the chances of the inconsistency with the increasing treatment effect having no impact.

3.7.2 Binary distributed outcome

To examine the corresponding behaviour of the overall odds ratio relative to that of the subgroup odds ratios a further series of simulations were undertaken. For each subject, a binary outcome was determined through the assignment of a response rate (binomial distribution) to each of the four treatment by factor combinations. The response rates were chosen such that both treatment and factor had an independent effect on outcome using the odds ratio model. That is, the odds ratio (T=2/T=1) for F=1 was identical to the odds ratio for F=2.

The range of values selected for the odds ratios for both the treatment and factor effects were 1, 1.25, 1.5, 2, 3 and 4, which generated a set of 36 basic combinations. In addition a reference odds was defined which fixed the odds for the T=1, F=1 combination of treatment and factor that enabled the odds to be defined for each of the 36 basic odds ratio combinations. The reference odds took the values 0.5 and 1, which resulted in 72 (36x2) different combinations for the odds. For a reference odds of 1, an odds ratio of 1.25 would correspond to a treatment (or factor) difference of around 5 percentage points (5.6%) whilst an odds ratio of 4 would correspond to a 30% difference. The effects 1.5, 2 and 3 would correspond to differences of 10%, 16.7% and 25% respectively. In this respect, these differences (in addition to no effect) were seen to represent the range of plausible effects in the clinical trial setting. The reference value of 0.5 was employed to ensure that some combinations contained odds which were both <1 and >1. The table with the highest odds was for the combination where both the odds ratios were 4 and the reference odds was 1. This gave the following 2x2 table of odds and response percentages: [1 = 50% (T=1 & F=1); 4 = 80% (T=1 & F=2); 4 = 80% (T=2 & F=1); and 16 = 94% (T=2 & F=2)]. For a given odds, binomial distributions were assigned to each treatment by factor combination using the relationship $\pi = \lambda/(1 + \lambda)$. The effect of sample size was also investigated using the three identical scenarios used for the Normally distributed outcomes. That is, 40, 200 and 1000 subjects. As a result, a total of 216 (36x2x3) simulations were performed.

In cases where either all or no subjects for a particular combination were assigned a success then 0.001 was added to both the number of successes and the number of failures to enable the odds to be defined. The rationale for selecting 0.001 was that with 500 being the maximum number of subjects per treatment group then a maximum odds would be (500.001/0.001) which would always be greater than (499/1). However this adjustment

was really only needed when the sample size was 40 – and in particular when the odds for a specific combination of treatment and factor was relatively high.

Two aspects of the odds model were examined. Firstly, an investigation of the potential for observing SP as the relative sizes of the treatment and factors effects, and sample size, varied. Secondly, an investigation of the potential for observing cases where the overall odds was either greater than, or smaller than, the ratio odds ratios for both subgroups. In this later case, the two different observations [>than and <than] have been reported separately to investigate the additional feature of unconditional odds ratio underestimation. In this respect, the expectation was that the respective proportions would be asymmetric in cases where the treatment and factor effects were not null. The results of the two series of simulations (2a and 2b) are presented in Tables 3.VII (reference odds = 1) and 3.VIII (reference odds = 0.5).

Table 3.VII. Simulation 2a: Binary outcomes: Percentage where overall treatment difference > or < treatment effect in both subgroups (% Simpson's paradox) with varying sample size ($\lambda_{11} = 1$)

Treatment OR (ψ_T)	Total sample size	% > or < and % SP	Factor OR (ψ_F)					
			1	1.25	1.5	2	3	4
1	40	>	**4.14**	4.10	4.56	5.62	7.74	9.12
		<	**4.10**	4.44	4.82	5.84	8.36	9.52
		SP	**(0.06)**	(0.14)	(0.16)	(0.20)	(0.42)	(0.64)
	200	>	**1.80**	2.38	3.52	5.96	8.88	9.58
		<	**1.80**	2.06	2.86	5.44	8.00	9.92
		SP	**(0.02)**		(0.02)	(0.18)	(0.68)	(1.16)
	1000	>	**0.70**	1.56	3.16	4.94	7.42	9.94
		<	**0.84**	1.32	3.42	4.92	8.20	10.26
		SP		(0.02)	(0.18)	(0.46)	(0.98)	(1.42)
1.25	40	>	3.48	**4.16**	4.26	4.70	6.98	8.36
		<	4.66	**5.38**	5.18	6.32	9.48	10.84
		SP	(0.04)	**(0.08)**	(0.14)	(0.16)	(0.48)	(0.74)
	200	>	1.86	**2.48**	3.22	4.24	6.36	7.30
		<	2.24	**2.64**	3.66	6.56	9.90	13.44
		SP		**(0.06)**	(0.02)	(0.22)	(0.56)	(0.72)
	1000	>	0.66	**1.56**	2.28	3.42	3.90	5.38
		<	0.82	**2.12**	4.10	6.88	12.56	16.78
		SP				(0.08)	(0.28)	(0.52)
1.5	40	>	3.14	3.40	**3.70**	5.44	5.84	7.14
		<	4.88	4.98	**5.54**	6.36	9.42	11.70
		SP	(0.06)	(0.12)	**(0.08)**	(0.28)	(0.46)	(0.46)
	200	>	1.26	2.08	**2.78**	3.54	4.94	5.40
		<	2.06	2.98	**4.02**	7.08	11.24	15.66
		SP			**(0.02)**	(0.16)	(0.40)	(0.54)
	1000	>	0.56	1.14	**1.84**	2.58	2.62	2.68
		<	0.96	2.44	**4.30**	9.40	17.04	22.44
		SP					(0.02)	(0.06)
2	40	>	2.70	3.18	2.94	**3.82**	5.38	6.34
		<	4.90	5.18	5.76	**7.52**	11.36	13.44
		SP	(0.06)	(0.04)	(0.08)	**(0.14)**	(0.22)	(0.60)
	200	>	1.78	1.56	2.16	**2.84**	3.56	3.86
		<	2.04	2.88	4.36	**8.20**	13.74	18.30
		SP					(0.08)	(0.06)
	1000	>	0.68	1.08	1.44	**1.52**	1.06	1.02
		<	1.14	2.54	5.98	**12.68**	23.44	32.46
		SP						
3	40	>	2.52	3.18	3.42	3.60	**5.02**	5.56
		<	5.22	6.54	7.18	8.54	**12.94**	15.68
		SP		(0.06)	(0.06)	(0.10)	**(0.08)**	(0.08)
	200	>	1.12	1.40	1.30	1.92	**2.16**	2.16
		<	2.86	4.12	5.66	9.26	**16.12**	23.08
		SP						
	1000	>	0.62	0.80	1.06	0.68	**0.38**	0.18
		<	1.12	3.30	7.32	16.80	**29.34**	42.72
		SP						
4	40	>	2.62	2.98	4.20	4.44	6.92	**7.56**
		<	6.14	6.46	6.58	8.58	13.12	**17.06**
		SP		(0.02)	(0.02)	(0.02)	(0.02)	**(0.14)**
	200	>	1.04	1.24	1.32	1.82	1.24	**1.46**
		<	3.40	3.68	5.60	10.42	18.40	**24.42**
		SP						
	1000	>	0.50	0.38	0.52	0.58	0.16	**0.16**
		<	1.30	3.00	7.18	17.20	33.80	**43.76**
		SP						

5000 simulated data sets with randomly assigned to F (p=0.5) but with balanced treatment groups

Table 3.VIII Simulation 2b: Binary outcomes: Percentage where overall treatment difference > or < treatment effect in both subgroups (% Simpson's paradox) with varying sample size ($\lambda_{11} = 0.5$)

Treatment OR (ψ_T)	Total sample size	% > or < and % SP	Factor OR (ψ_F)					
			1	1.25	1.5	2	3	4
1	40	>	**3.48**	3.88	4.36	6.26	9.10	9.96
		<	**3.74**	4.28	4.20	6.16	9.54	10.50
		SP	**(0.06)**	(0.04)	(0.12)	(0.24)	(0.56)	(0.86)
	200	>	**1.98**	2.34	3.62	4.80	8.48	10.78
		<	**1.76**	2.32	3.20	5.30	8.22	10.78
		SP		(0.02)	(0.06)	(0.30)	(0.84)	(1.40)
	1000	>	**0.90**	1.84	3.46	4.92	8.36	10.14
		<	**1.08**	2.08	2.92	5.76	8.72	8.74
		SP		(0.06)	(0.08)	(0.34)	(1.14)	(1.20)
1.25	40	>	3.82	**3.76**	3.74	5.62	7.96	8.72
		<	4.14	**4.32**	5.46	6.86	10.04	12.60
		SP	(0.02)	**(0.04)**	(0.10)	(0.18)	(0.58)	(1.16)
	200	>	1.90	**1.74**	2.56	4.34	6.16	8.30
		<	1.84	**2.46**	3.48	6.46	10.04	14.34
		SP				(0.18)	(0.50)	(1.10)
	1000	>	0.80	**1.44**	2.44	3.42	4.46	4.80
		<	0.76	**1.94**	3.42	8.24	13.48	18.64
		SP			(0.02)	(0.12)	(0.32)	(0.36)
1.5	40	>	3.38	3.68	**4.34**	5.30	6.38	7.42
		<	3.66	4.98	**6.30**	6.70	9.92	12.70
		SP	(0.04)	(0.06)	**(0.20)**	(0.30)	(0.40)	(0.80)
	200	>	1.50	1.78	**3.02**	4.04	4.76	5.30
		<	2.28	2.94	**4.58**	7.32	13.04	17.70
		SP		(0.02)	**(0.04)**	(0.14)	(0.26)	(0.46)
	1000	>	0.78	1.42	**1.96**	2.76	2.26	2.18
		<	0.76	2.48	**4.78**	9.36	20.22	27.12
		SP				(0.02)		(0.04)
2	40	>	2.56	3.10	3.34	**4.52**	5.38	6.28
		<	5.28	5.54	6.68	**8.62**	12.40	15.82
		SP		(0.08)	(0.08)	**(0.16)**	(0.22)	(0.54)
	200	>	1.44	1.80	2.42	**3.00**	3.38	3.28
		<	2.24	3.02	4.84	**8.84**	17.26	22.16
		SP				**(0.02)**	(0.08)	(0.12)
	1000	>	0.82	1.30	1.56	**1.42**	0.96	0.48
		<	0.84	3.16	6.34	**12.86**	28.88	39.52
		SP						
3	40	>	2.24	2.68	2.60	3.62	**3.56**	4.22
		<	6.48	6.34	7.90	10.06	**14.58**	18.20
		SP	(0.04)	(0.04)	(0.08)		**(0.12)**	(0.22)
	200	>	1.08	1.10	1.74	2.02	**1.92**	1.38
		<	2.40	3.50	5.76	11.48	**21.14**	28.82
		SP						
	1000	>	0.54	0.68	0.56	0.68	**0.22**	0.02
		<	1.44	3.52	7.38	18.66	**40.22**	54.44
		SP						
4	40	>	2.38	2.16	2.26	2.78	2.92	**3.82**
		<	7.34	7.24	8.78	11.64	15.60	**20.82**
		SP			(0.02)	(0.04)	(0.04)	**(0.16)**
	200	>	1.10	1.00	1.24	1.12	1.36	**0.92**
		<	3.52	4.16	6.64	13.28	23.54	**32.22**
		SP						
	1000	>	0.44	0.44	0.70	0.24	0.10	
		<	1.38	4.06	8.86	22.94	45.48	**61.90**
		SP						

5000 simulated data sets with randomly assigned to F (p=0.5) but with balanced treatment groups

Simpson's paradox

The incidence of SP was again low under the range of conditions selected. The maximum incidence (1.42%) occurred when there was no treatment difference (odds ratio =1), the reference odds was 1 and the factor effect (4) was at it greatest, although in contrast to the Normal distributed outcome case the sample size was at the highest limit of the conditions investigated (N=1000). When the reference odds was 0.5, the maximum incidence (1.40%) also occurred when the treatment difference was zero and the factor effect was four – although in this case the sample size was 200.

As expected, for a given treatment difference (table row), the incidence of SP increased with increasing size of factor effect while the incidence decreased with increasing size of treatment effect when the factor difference was fixed (table column). SP was not observed for many of the combinations - in particular, when the sample size was 1000 and the treatment odds ratio was >1 .5.

For a given combination of treatment and factor effect, the incidence of SP decreased with increasing sample size when the treatment odds ratio was >1. However when the treatment odds ratio was one, the incidence of SP actually increased with increasing sample size at the higher factor effects. For instance, when the reference odds was one and the factor effect was four the incidence of SP increased from 0.64% to 1.16% to 1.42% for the sample sizes 40, 200 and 1000 respectively. Two additional simulations for 5000 and 25000 subjects confirmed this trend with incidences of 1.42% and 1.50% respectively. Furthermore, when the reference odds was 0.5, sample sizes of 5000 and 25000 produced incidences of 1.88% and 1.74% respectively - both higher than those generated for sample sizes of 40 (0.86%), 200 (1.40%) and 1000 (1.20%). At these higher sample sizes the observed odds ratios tended to be close to one so, although there was a

higher proportion of reversals, the differences between the overall and subgroups odds ratios were less pronounced than those that typically occurred with the lower sample sizes. This would be expected as the precision of the cell estimates increase with increasing sample size.

Regarding the pattern of SP when the treatment and factors odds ratios were identical (table diagonal bolded), it is worth noting that unlike the Normal case, the variability does not remain constant but instead changes dependent upon the observed response proportions. When the sample size was 1000, the incidence of SP was zero (for both 1 and 0.5 reference odds) for all cases were the treatment and factors odds ratio were identical. Similarly there were few occurrences of SP when the sample size was 200. When the sample size was 40, it was difficult to discern a trend although perhaps there was a tendency for an increased incidence of SP as the treatment and factor odds ratios increased.

When the treatment odds ratio was greater than one, the subgroups were most likely to show the correct direction of the treatment difference – rather than the unconditional treatment odds ratio - when the factor effect was large and so was the sample size. For example, when the treatment odds ratio was 1.25, the factor odds ratio was 4 and the reference odds was 0.5, then 12 of the 18 cases of SP observed showed the correct direction of the treatment difference in the subgroups when the sample size was 1000. However, as per the Normal case, when both the sample size and factor effect decreased, the assignment of correct direction became more balanced between the subgroups and the unconditional treatment difference.

Overall effect greater than or less than both subgroups

As discussed earlier, when considering the more general inconsistency, it was important to distinguish the two conditions (greater than both subgroups [>than] and less than both subgroups [<than]) in this sub-section as a result of the known underestimation of the unconditional non-unity odds ratio in the balanced non-unity factor case. Indeed, as will be described below, marked asymmetry of effect was observed, justifying this approach for the simulations.

As with SP, the incidence of both conditions (>than and <than) increased with increasing factor effect (table row). However as the treatment effect was increased, the gradient of the increased incidence was much more apparent for the <than condition than for the >than. Indeed when the factor effect was 4 and the sample size was high there was an indication that the incidence for the >than condition was actually beginning to decrease. When the factor and treatment effects were both 4 and the reference odds was 0.5, then at a sample size of 1000 there were no occurrences of >than while the incidence of <than was 61.90%. That is, the unconditional odds ratio was less than both subgroup odds ratios for over 60% of trials and was not greater than both subgroup odds ratios a single time. When the reference odds was 1 the respective proportions were 43.76% and 0.16%. Interestingly in contrast to the Normal case, increasing the sample size when the factor and treatment effects were high only served to increase the total incidence (>than + <than) whilst increasing the difference between the two combinations.

For a fixed factor effect and sample size (table column), increasing the treatment effect reduced the incidence of >than but increased the corresponding incidence of <than. In general the overall incidence (>than + <then) increased also, although increase was more marked at the higher factor effects investigated.

Increasing the sample size for a given combination of treatment and factor effects reduced the incidence of >than but increased the incidence of <than when the factor effect was large. However when the factor effect was small the incidences of both inconsistencies decreased with increasing sample size.

When the treatment and factors effects were identical (table diagonal), the overall incidence of the inconsistency (>than + <than) increased. In particular the <than incidence clearly increased although the >than incidence appeared to increase then decrease in most cases.

When comparing the two different reference odds, 1 (Tables 3.VII) versus 0.5 (Table 3.VIII), it was notable that the difference between the incidences (<than minus >than) for a particular combination of factor, treatment and sample size was smallest for the reference odds of 1. This would be expected since in this case the odds ratios would be closer to the relative risk and as such greater symmetry would be expected.

3.8 DISCUSSION

Observed imbalance is an integral part of randomised studies and it is this imbalance which leads to the application of statistical techniques that account for the resulting variability. The chapter focuses upon the impact of factor imbalance on observed data. In order not to detract from the main focus of this chapter, all of the examples have been deliberately constructed so as to avoid interaction terms. That is not to say that interactions are not important – although ones of a quantitative nature are sometimes a simple artefact of the chosen scale of measurement (Gail and Simon, 1985) - rather that the potential impact of simple imbalance in a RCT should not be understated.

Simpson's paradox is a useful starting point when considering the impact of factor imbalance in relation to the effect on estimated treatment differences since two key aspects of the data are present. These are a factor that has an independent influence on outcome and imbalance in the distribution of this factor between the treatments. Indeed these are the only two data features required for SP. (Although SP could be generated from a model that contained a treatment by factor interaction, it has been shown that this is, by no means, a necessary condition.) As such, SP may simply reflect imbalance between treatment groups with respect to a factor that in general has a greater impact on outcome that the differential effect of the treatments being compared.

Although it has been shown that randomisation provides a theoretical basis for avoiding SP when subgroups are defined appropriately, the simulations have shown that SP can still occur in a RCT when the conditions are right. For the simple additive model selected, the most favourable conditions to observe SP were an equivalence (or non-inferiority) design with no treatment difference where an independent factor existed that had a large influence on outcome. That being said, SP also occurred when the treatment difference was not unity. In these cases, when SP was observed it was more likely that the estimated treatment differences in the subgroups showed the true direction of the treatment effect compared to the corresponding unconditional estimate. This was particularly the case when both the factor effect and sample size were set towards the top end of their respective ranges. In terms of choosing between estimates, since the observed treatment differences are likely to be on the small side if SP is observed, the overall interpretation is in fact unlikely to be influenced by the choice. This is particularly so with modest to large sample sizes. That being said, if the factor is based on post-randomisation data (such as compliance) then it is clear that the unconditional estimate is the most appropriate one to

select. In contrast, if the factor is a design feature - that is, *a priori* stratification by the factor has been undertaken - then the conditional estimate is the most appropriate. For other pre-randomisation factors that are known *a priori* to influence outcome then conditional estimates are again appropriate. The more challenging situation *is post hoc* stratification for factors whose impact on outcome is unknown. These tend to form part of an exploratory analysis and as such a conditional estimate based on these factors should not be primary.

Notwithstanding, SP is likely to be a rare event in the randomised clinical trial setting and when it does occur it is most likely that it would go unnoticed or unreported. As the simulations demonstrate, the results tend not to be too dramatic and the overall unconditional treatment difference is likely to be small. Under the range of conditions investigated in the simulations, the highest incidences were 8.88% for Normally distributed outcomes and 1.42% for binary distributed outcomes. Perhaps a reason for SP going unnoticed is that frequently just the estimated unconditional and conditional treatment differences are presented and not the individual subgroup data. In this respect a change in sign between the two estimates may indicate a potential SP. With the growth in non-inferiority studies and the current regulatory interest in subgroups it is indeed possible that cases of SP may come to light in the future even in the RCT setting. (Superiority studies of unpromising test treatments may also provide a hunting ground for such effects - although negative superiority studies are seldom formally published!)

Using SP as a vehicle, it has also been demonstrated that compared to the overall treatment difference, it is a misconception that an increase in the size of estimated treatment difference in one subgroup must lead to a corresponding reduction in the other. For Normally distributed outcomes, the incidence of either observation (<than or >than)

reached 50% under the conditions investigated, with symmetry of the proportions, <than and >than, as expected. However with the odds model it is much more likely that the difference between treatments will be larger in both subgroups than overall, rather than *vice versa*. Under the conditions investigated in the simulations, the incidence reached 60% for the observation that the odds ratio was larger in both subgroups compared to overall - in contrast there was no occurrence of the reverse phenomenon for the identical condition. These observations serve to highlight another potential pitfall in the notoriously difficult task of interpreting subgroup analyses. In particular one should not assume that if the odds ratio in one subgroup is observed to be larger than the unconditional estimate then the odds ratio in the complement subgroup will necessarily be smaller.

Overall the series of simulations illustrate that when randomisation is employed to assign treatments, there is a complex interplay between the sometimes-opposing effects of treatment, factor, sample size and underlying variability that can sometimes produce apparent inconsistencies of small or modest effect. Indeed it has been shown that increasing the sample size does not necessarily reduce the chances of these inconsistencies occurring and under some conditions the chances of an inconsistency are actually increased. Furthermore the impact of changing a condition does not necessarily lead to the same effect on SP as it does on the more general inconsistency (<than and >than). For instance, in the Normally distributed outcome case, increasing the treatment difference reduces the incidence of SP but has no impact on the general inconsistency.

The increased interest now being given to particular subjects subgroups by the regulatory authorities suggests that emphasis is moving away from simply reporting average treatments effects towards providing more specific treatment effect information for

particular subgroups. Relevant guidelines include the ICH documents (E3; E5; E7; E9; E11) and the CPMP Points to Consider (multiplicity issues in clinical trials; adjustment for baseline covariates). Hand (1994) states that *much statistical analysis and design is misdirected*. That is, it is important to establish which questions to ask and to answer these appropriately. In particular one should avoid providing the *right answer to the wrong question*. Lane and Nelder (1982) describe model selection, model fitting and prediction in the context of generalised linear models. They describe prediction not in terms of future events rather a case of *what would have happened in the experiment if other conditions had prevailed*. In this context Nelder (1994a) explains how parameter estimates can be used to construct *quantities of interest* using weighting schemes that are appropriate to the questions posed. Indeed, in the reporting of clinical trial data in the pharmaceutical and biotechnology industries little thought is currently given to prediction and almost all reporting activity is directed exclusively to describing what actually happened in a particular study or series of studies. For clinical trials, prediction could easily relate to the standardisation of results to the overall diseased population where the proportions of subjects with defined characteristics - such as sex and race - are obtained from national health statistics, and providing estimates of the treatment difference in specific subgroups is an important first step. But this approach is not without its own set of problems. As highlighted by Senn (1997), subjects can not be assigned at random to their characteristics. For instance, the females included in a particular trial may not be representative of females in the disease population in general - as highlighted earlier in Chapter Two. In this case, the estimated difference between the male and female subgroups may actually represent the exclusion of an important subgroup of subjects - such as pregnant women or those of child bearing potential - rather than a difference between the sexes *per se*. This point is particularly relevant to prediction in that one has to ensure that the data warrant generalisation to the broader population.

Nevertheless, it makes sense to use the parameter estimates from confirmatory clinical trials, where restrictions to enrolment are becoming less strict, to construct quantities of interest in an effort to provide specific answers to targeted questions regarding the impact of new treatments. Indeed, one such question might be directed towards determining whether treatment differences are uniform across specific subgroups, and this is the theme of the following chapter (Chapter Four). That is, the investigation of quantities of interest for the interaction between treatments and subgroups.

APPENDIX A. FORMULA FOR UNCONDITIONAL ODDS RATIO

A simple formula for calculating the unconditional odds ratio (ψ) from a combination of the four separate odds (λ_{ij}) can be derived if it is assumed that the number of subjects in each treatment (i=1,2) by factor (j=1,2) combination is identical. Consider Table 3.A.I which gives the numerators and denominators for the 4 treatment by factor combinations and for the two treatments combined across the two levels of factor F.

Table 3A.I. Observed outcome proportions (x_{ij} / n_{ij}) for a two treatment, two factor design

	Factor F=1	Factor F=2	Total
Treatment T=1	x_{11} / n_{11}	x_{12} / n_{12}	$(x_{11} + x_{12}) / (n_{11} + n_{12})$
Treatment T=2	x_{21} / n_{21}	x_{22} / n_{22}	$(x_{21} + x_{22}) / (n_{21} + n_{22})$

Using the notation above, the unconditional odds ratio (ψ) is defined as:

$$(x_{11} + x_{12})\,[(n_{21} - x_{21}) + (n_{22} - x_{22})] \,/\, (x_{21} + x_{22})\,[(n_{11} - x_{11}) + (n_{12} - x_{12})] \qquad (A1)$$

Now, the odds for each treatment by factor combination is given by $\lambda_{ij} = x_{ij} / (n_{ij} - x_{ij})$ and it follows that $x_{ij} = n_{ij}\,\lambda_{ij} / (1 + \lambda_{ij})$ and $(n_{ij} - x_{ij}) = n_{ij} / (1 + \lambda_{ij})$.

Substituting the specific terms, x_{ij} and $(n_{ij} - x_{ij})$ into (A1), for each of the four combinations treatment by factor combinations (i =1, 2; j=1, 2) and further assuming complete balance ($n_{11} = n_{12} = n_{21} = n_{22} = n$), gives ψ which is independent of n:

$$\psi = (\lambda_{11} + \lambda_{12} + 2\lambda_{11}\lambda_{12})(2 + \lambda_{21} + \lambda_{22}) / (\lambda_{21} + \lambda_{22} + 2\lambda_{21}\lambda_{22})(2 + \lambda_{11} + \lambda_{12})$$

APPENDIX B. REDEFINITION OF BALANCE FOR THE ODDS MODEL

Proof that a redefinition of the concept of balance for the odds model enables consistency to be achieved between the results of overall and subgroup summaries.

Using the notation given previously in Table 3.A.I it is possible to show the result more formally. If the proportion of failures is identical in each group then:

$$(n_{11} - x_{11}) / [(n_{11} - x_{11}) + (n_{12} - x_{12})] = (n_{21} - x_{21}) / [(n_{21} - x_{21}) + (n_{22} - x_{22})] \qquad \text{(B1)}$$

which can be re-arranged to give:

$$(n_{11} - x_{11}) = (n_{12} - x_{12})(n_{21} - x_{21}) / (n_{22} - x_{22}) \qquad \text{(B2)}$$

Further, if the odds ratio is identical for each level of factor F then:

$$x_{11}(n_{21} - x_{21}) / x_{21}(n_{11} - x_{11}) = x_{12}(n_{22} - x_{22}) / x_{22}(n_{12} - x_{12}) \quad \text{(B3)}$$

Substituting (B2) into (B3) gives:

$$x_{11} = x_{12}\ x_{21} / x_{22} \qquad \text{(B4)}$$

Now, from Appendix A (A1), the unconditional odds ratio is defined as:

$$(x_{11} + x_{12})[(n_{21} - x_{21}) + (n_{22} - x_{22})] / (x_{21} + x_{22})[(n_{11} - x_{11}) + (n_{12} - x_{12})]$$

Substituting (B2) and (B4) into (A1) gives:

$x_{12}(n_{22} - x_{22}) / x_{22}(n_{12} - x_{12})$

which is identical to the odds ratio when F=2.

CHAPTER FOUR: DIFFERENT DIFFERENCES

Johnny Hammer
Had a terrible ss..ss..ss..ss..ss..ss..stammer
He could hardly s..s..say a word
And so they gave him medicinal compound
Now he's seen (but never 'eard)!

4.1 INTRODUCTION

In general, an interaction can be described as the case where the effect of one factor, depends upon the level of another factor (Pocock, 1983), and in the clinical trial setting, it is the impact of factors on the relative effects of treatment that is of interest. In this respect, different levels of a factor form mutually exclusive subgroups and treatment by factor interactions equate to differences between the treatment differences amongst subgroups – the presence of which may have important implications for the treatment of patients. This point has not been lost on US regulatory authorities who require pharmaceutical companies to determine the level of support provided by their data for the proposed dose schedule of a new treatment across specific subgroups (FDA, 1988). Indeed when taken alongside developments in the area of genetics, interactions will undoubtedly remain an area of great interest to researchers and regulators alike.

There are many challenges to be faced when evaluating interactions in clinical trials. Firstly, the relative power of the traditional hypothesis test approach is low. As introduced in Chapter Two, the standard error for the estimate of the interaction parameter is not only larger than the standard error for the overall treatment difference but it is also larger than the standard error for the treatment difference in each of the subgroups. As such, a lack of statistical significance may not provide robust support for the conclusion that no interaction exists. Second, the size of effect that constitutes a clinically relevant

difference is not well established for interaction parameters and while the distinction between a qualitative and a quantitative interaction (first introduced in Chapter Two) may have important treatment implications, this information is not fully captured by the value of the interaction parameter. Third, the interpretation of unexpectedly large interactions is difficult – particularly those of a qualitative nature – and as a result there is a need to place such findings in context, based on our current state of knowledge.

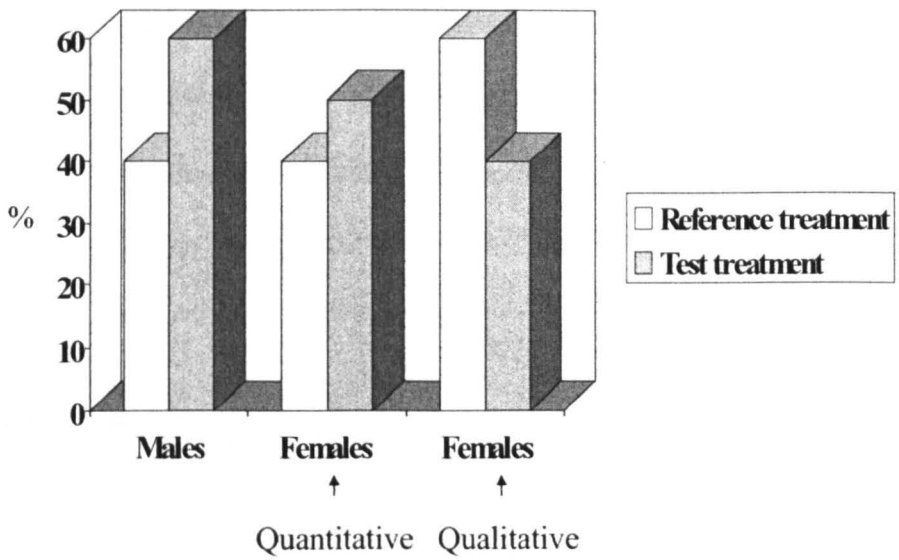
The aim of this chapter is to explore each of these aforementioned challenges in more detail and, where appropriate, to propose alternative approaches to the evaluation and interpretation of treatment by factor interactions. Section 4.2 presents a general overview of treatment by factor interactions. The impact of the scale of measurement on interaction interpretation is described and the relative precision of the estimate of the interaction parameter in relation to the corresponding main and subgroup effects is investigated. The important distinction between qualitative and quantitative interactions is discussed in detail and the statistical tests that have been developed in this area are reviewed. In section 4.3, regulatory guidance is reviewed, and then in Section 4.4, a framework is developed to consolidate methods of evaluation based on the value of the interaction parameter with those aimed at distinguishing between interactions of a quantitative and qualitative nature. Different approaches to determining what size of effect might be considered clinically relevant are also discussed. In Section 4.5, a simple Bayesian approach to the evaluation of interactions is considered while in Section 4.6 an example is presented to illustrate this Bayesian approach and how it could be applied to interpret some real clinical trial data in practice. Finally in Section 4.7, the findings of this chapter are discussed.

4.2 GENERAL CONSIDERATIONS

4.2.1 Quantitative versus qualitative interactions, and data transformations

As introduced in Chapter Two (Section 2.4.1), Peto (1982) distinguishes between two types of interaction - qualitative and quantitative. A quantitative interaction is one where the true direction of the treatment difference is the same in the subgroups but the magnitude of the difference is different. This is illustrated using hypothetical data in Figure 4.1 by comparing the relative treatment effect for the subgroup of males with that of the subgroup of females as represented by the two middle columns. In this case, the response percentage is 20% higher with Test treatment compared with Reference in the subgroup of males whereas in the subgroup of females the response rate is still higher with Test treatment, but the difference is now only 10%.

Figure 4.1. Quantitative versus qualitative interaction



In contrast, a qualitative interaction is one where the true direction of the treatment difference is different amongst subgroups. In Figure 4.1, this is illustrated by comparing the males with the subgroup of females represented by the final two columns where the

response percentage is actually 20% higher with the Reference treatment. (Note that Gail and Simon (1985) substitute the terms *crossover* and *non-crossover* for qualitative and quantitative interactions respectively - and use the term quantitative to describe interactions of any kind. To avoid confusion the Peto terminology will be used throughout this chapter and research thesis.)

Now, in cases where a factor is known to influence outcome, Peto expects *a priori* that a quantitative interaction with treatment will exist - regardless of whether one is observed or not. In a similar vein, Gail and Simon (1985) expect quantitative interactions *because there is usually no self-evidently appropriate scale of response measurement* but highlight that these may not be clinically relevant. The rationale for this view is that a quantitative interaction is sometimes a simple artefact of the chosen scale of measurement (Hand, 1994; Gail and Simon, 1985). For instance, as illustrated in Chapter Two, Table 3.I, what may appear to be an interaction when evaluating proportions can disappear when the data are summarised in terms of odds (or *vice versa*). Similarly, as noted also in Chapter Two, Gail and Simon (1985) illustrate how the logarithmic transformation of continuous data can lead to the disappearance of an apparent interaction. (That is, $y = e^{\alpha} e^{\beta}$ transformed to $\log_e y = \alpha + \beta$.)

An interaction of the qualitative type cannot be transformed to the quantitative type (or *vice versa*) when moving between the difference in proportions, relative risk and odds ratio (OR). (This is easily proved, since if $p_t - p_r > 0$ then $\frac{p_t}{p_r} > 1$, and since the OR can be written $\frac{p_t}{p_r}\left(\frac{1-p_r}{1-p_t}\right)$, then since also $1 - p_t < 1 - p_r$, the OR must also be greater than one.) However in the case of continuous data, some data transformations do actually have

the potential to transform a quantitative interaction to a qualitative one (or *vice versa*). Consider Table 4.I that illustrates the common "change from baseline" endpoint and compare this with a transformation to "percentage change from baseline".

Table 4.I: Absolute and percentage change from baseline comparison

		Factor F=1				Factor F=2			
		BL	FU	FU-BL	%	BL	FU	FU-BL	%
Test	Subjects	1	2	+1	+100	2	1	-1	-50
	Mean			+1	+100			-1	-50
Reference	Subjects	1	2	+1	+100	2	1	-1	-50
		20	18	-2	-10	20	18	-2	-10
	Mean			-0.5	+45			-1.5	-30
Test - Reference				+1.5	+55			+0.5	-20

BL=Baseline; FU=Follow-up

In this simple example, with just six subjects and a factor with two levels, the differences between treatments (Test - Reference) with regard to the mean absolute change from baseline (to a treated follow-up visit) are +1.5 for factor F=1 and +0.5 for factor F=2, indicating a quantitative interaction. However, the difference between treatments with regard to the mean percentage change from baseline is +55% for factor F=1 but -20% for factor F=2, indicating a qualitative interaction. In this illustration, the paradox is due to the wide range of baseline scores between subjects taken alongside the small range of absolute changes from baseline. That is, the impact of the different weighting of the changes from baseline in each case.

Now, Peto is highly sceptical about the presence of qualitative interactions unless *good prior reasons* exist. However in equivalence or non-inferiority trials there may be a case for being much less sceptical. For example in septicaemia, a new anti-infective could be more effective than a reference treatment in subjects with *Escherichia coli* bacteria, but less effective in subjects with *Staphylococcus aureus* bacteria, giving rise to a qualitative interaction. Indeed the very nature of anti-infective treatments is that they tend to have

varying degrees of success against different bacteria. In many cases, according to Pan and Wolfe (1997), *a slight qualitative interaction is both natural and expected* when comparing a test treatment to an effective reference treatment. (They quote the example of non-prescription painkillers, where no treatment is universally more effective than the others over the full range of pain indications.) Clearly the most likely conditions for observing a qualitative interaction are when active treatments with different modes of action are being compared, and where the average effects in a wide population are broadly similar. In contrast, qualitative interactions are much less scientifically plausible in placebo controlled studies unless an active treatment has a detrimental effect in a subgroup of subjects.

4.2.2 Relative precision and power

It is widely accepted that traditional hypothesis tests directed towards interaction parameters have low power relative to tests of overall treatment differences - a view based on the relative precision of the estimates of the corresponding parameters. Indeed the CPMP expresses the following concern: *Tests for interaction usually lack statistical power and the absence of statistical evidence of an interaction is not evidence that there is no clinically relevant interaction* (Points to Consider on adjustment for baseline covariates [CPMP/EWP/2863/99, 2002]). Now, to understand the basis for such a view, the simple case of a continuous outcome variable will be presented in terms of the relative precision of the estimates of specific parameters.

Consider a randomised and controlled clinical trial (RCT) with two treatment groups – test (t) and reference (r) – and a baseline factor (F) with two levels such that subjects belong to one of two mutually exclusive subgroups (F=1, 2). For each treatment by factor combination let the mean value for some continuous outcome be μ_{ij} with corresponding

estimate $\bar{x}_{ij}$. Also, let the mean values be μ_t and μ_r for the test and reference treatments respectively when factor F is ignored and let δ be the corresponding overall treatment difference, $\mu_t - \mu_r$.

		Factor	
		F= 1	F= 2
Treatment	T= t	μ_{t1}	μ_{t2}
	T= r	μ_{r1}	μ_{r2}

The interaction (Ω) between the treatment and the factor can then be written as (μ_{t1} - μ_{r1}) - (μ_{t2} - μ_{r2}). That is, the difference between test and reference treatments for F=1 minus the corresponding difference for F=2. The estimate of the interaction parameter ($\hat{\Omega}$) is then $(\bar{x}_{t1} - \bar{x}_{r1}) - (\bar{x}_{t2} - \bar{x}_{r2})$.

Now, as introduced in Chapter Two, if the variance of the overall treatment difference, $Var(\bar{x}_t - \bar{x}_r)$, is v, in some arbitrary units, then the variance of the treatment difference, $Var(\bar{x}_{tj} - \bar{x}_{rj})$, within each of two subgroups of equal size is $2v$, while the variance of the interaction term, $Var(\hat{\Omega})$, is $4v$ (Peto, 1982). Although Peto (1982) does not provide further detail it is straightforward to prove the relationship as follows:

Let the variance of the overall treatment difference be $Var(\bar{x}_t - \bar{x}_r) = \upsilon^2 / 2n = v$. Now, the variances of the treatment differences in each subgroup will be equal, but instead of being based on $2n$ observations, these are based on n observations each leading to variances of $Var(\bar{x}_{tj} - \bar{x}_{rj}) = \upsilon^2 / n = 2v$. It follows that the variance of interaction is the variance of the difference between the treatment differences from the two subgroups; that

is, $Var\{(\bar{x}_{t1} - \bar{x}_{r1}) - (\bar{x}_{t2} - \bar{x}_{r2})\}$. As such, $Var(\hat{\Omega}) = Var(\bar{x}_{t1} - \bar{x}_{r1}) + Var(\bar{x}_{t2} - \bar{x}_{r2})$; that is, $2\upsilon^2 / n = 4v$.

For completeness, since the overall weighted treatment difference – which gives equal weight to each treatment by factor combination - can be written:

$$\Lambda = \left(\frac{\mu_{t1} + \mu_{t2}}{2}\right) - \left(\frac{\mu_{r1} + \mu_{r2}}{2}\right) \qquad (1),$$

it follows that the variance of the estimate, $Var(\hat{\Lambda})$, is also v. That is,

$$Var\left\{\left(\frac{\bar{x}_{t1} - \bar{x}_{r1}}{2}\right) + \left(\frac{\bar{x}_{t2} - \bar{x}_{r2}}{2}\right)\right\} = \frac{1}{2^2} Var(\bar{x}_{t1} - \bar{x}_{r1}) + \frac{1}{2^2} Var(\bar{x}_{t2} - \bar{x}_{r2}) = v,$$

It is then simple to see, that in this case where the numbers of subjects in each subgroup are equal, $4Var(\hat{\delta}) = 4Var(\hat{\Lambda}) = 2Var(\bar{x}_{tj} - \bar{x}_{rj}) = Var(\hat{\Omega})$. Hence, the precision of the interaction (defined as the inverse of the variance) is half that of the subgroup treatment difference and a quarter that of the overall treatment difference.

At this point it is worth highlighting why the $Var(\hat{\Lambda}) \neq Var(\hat{\Omega})$ given that both are based on contrasts involving the parameters $\mu_{t1}, \mu_{t1}, \mu_{t1}, \mu_{t1}$, but with varying signs. The difference arises because, as shown in (1), the parameters, μ_{ij}, are all multiplied by a factor of $^1/_2$ for $E(\hat{\Lambda})$ and since in general $Var(bx) = b^2 Var(x)$ this leads to a multiplication factor of $^1/_4$ when estimating the corresponding variance, $Var(\hat{\Lambda})$.

Now, (1) can be re-arranged:

$$2\Lambda = \mu_{t1} - \mu_{r1} + \mu_{t2} - \mu_{r2} \qquad (2),$$

while the interaction, Ω, can also be re-arranged:

$$\Omega = \mu_{t1} - \mu_{r1} - \mu_{t2} + \mu_{r2} \qquad (3).$$

Combining (2) and (3) gives:

$$\Omega = 2(\Lambda - \mu_{t2} + \mu_{r2})$$

which can be re-arranged:

$$\Omega = -2(\mu_{t2} - \mu_{r2}) + 2\Lambda \qquad (4)$$

Now, $Var(\hat{\Omega})$ can also be written in the form:

$$Var(-2(\bar{x}_{t2} - \bar{x}_{r2}) + 2\hat{\Lambda}) = 4Var(\bar{x}_{t2} - \bar{x}_{r2}) + 4Var(\hat{\Lambda}) - 8Cov(\bar{x}_{t2} - \bar{x}_{r2}, \hat{\Lambda}),$$

from which it follows that the covariance, $Cov(\bar{x}_{t2} - \bar{x}_{r2}, \hat{\Lambda})$, is also ν. This leads to a correlation coefficient, $\rho(\bar{x}_{tj} - \bar{x}_{rj}, \hat{\Lambda}) = \dfrac{Cov(\bar{x}_{tj} - \bar{x}_{rj}, \hat{\Lambda})}{\sqrt{Var(\bar{x}_{tj} - \bar{x}_{rj}).Var(\hat{\Lambda})}}$, of $\frac{1}{\sqrt{2}}$ which illustrates the positive relationship between the treatment difference in the subgroup and overall weighted treatment difference.

If the numbers of subjects in each subgroup are unequal then the relative precision of the estimates differ but the general findings remain valid. For instance if one subgroup (F=1, say) were to contain three times as many subjects as the complement subgroup (F=2) then the variances would be $1\frac{1}{3}\nu$ and 4ν respectively, while the variance of the interaction would be the sum of the two - that is, $5\frac{1}{3}\nu$. The variance of the overall weighted difference, $Var(\hat{\Lambda})$, is $1\frac{1}{3}\nu$ which is now larger than the corresponding unweighted difference. The covariance terms for each subgroup also differ. For F=1 the covariance is now $\frac{2}{3}\nu$ with a correlation coefficient of $\frac{1}{2}$, whereas the corresponding values for F=2 are 2ν and $\frac{\sqrt{3}}{2}$ respectively. (Note this implies a stronger correlation between the weighted treatment difference and the subgroup with fewer subjects compared with the subgroup with more subjects.)

In the more general case of multiple treatments and multiple subgroups (from a single factor) then the model may take the form $\mu_{ij} = \alpha + \tau_i + \phi_j + (\tau\phi)_{ij}$ leading to (i-1)(j-1) interaction parameters to estimate. (For instance three treatments and three subgroups generate 4 parameters to estimate and very quickly the overall picture becomes much more difficult to interpret.) However the focus in this chapter will be on the simple two treatment, two subgroup case where one interaction parameter is of interest.

4.2.3 Influential work in the area of interactions

Gail and Simon (1985) in their seminal paper introduced the first test aimed at distinguishing between qualitative and quantitative interactions. Their likelihood ratio (LR) test - based here on the notation of Piantadosi and Gail (1993) - considers independent and normally distributed estimates (D_i) of the treatment differences in j=1, 2,..., J subgroups (with mean δ_j and known variance σ_j^2). The two sided null hypothesis takes the form $H_{02} : \Delta \in 0^+ \cup 0^-$ where Δ is the vector of parameters (δ_j), and 0^+ and 0^- represent two orthants such that all the parameters contained therein are either ≥ 0 (0^+) or ≤ 0 (0^-). (Note that these two orthants include zero treatment effect and that a qualitative interaction is indicated in the remaining 2^{J-1} orthants.) In this respect the null hypothesis (of no qualitative interaction) considers the case where either all the treatment differences (δ_j) are non negative or all are non positive, and is rejected if both

$$Q^- \equiv \sum \left\{ \left(D_j^2 / \sigma_j^2 \right) I\left(D_j > 0 \right) \right\} > C_{2\alpha}$$

$$Q^+ \equiv \sum \left\{ \left(D_j^2 / \sigma_j^2 \right) I\left(D_j < 0 \right) \right\} > C_{2\alpha}$$

where $C_{2\alpha}$ represents the critical value (α level) for the test and I is an indicator variable which equals one if the condition is true, otherwise zero. Gail and Simon (1985) also present a one sided null hypothesis ($H_{01} : \Delta \in 0^+$). In this case the null hypothesis is that

test treatment is at least as good as the reference treatment in all subgroups, which is rejected if:

$$Q^{+} > C_{1\alpha},$$

where $C_{1\alpha} > C_{2\alpha}$. In practice, consistent estimates of the unknown σ_j^2 (that is, s_j^2) are used in the formulae and result in valid asymptotic significance tests.

As a simple alternative to the LR test, Piantadosi and Gail (1993) formally introduced the standardised range (SR) test that rejects H_{02} at the α level if both the following conditions are met:

$$\max\{D_j / \sigma_j\} > C'_{2\alpha}$$

$$\min\{D_j / \sigma_j\} < -C'_{2\alpha}$$

Similarly H_{01} is rejected if

$$\min\{D_j / \sigma_j\} < -C'_{1\alpha}.$$

(Note that Piantadosi and Gail actually credit Robert Tarone with devising the SR test.) Again $C'_{1\alpha}$ and $C'_{2\alpha}$ represents the respective one and two sided (test specific) critical values at the α level.

Piantadosi and Gail (1993) compared the LR and SR tests with regard to power and concluded that for two or three subgroups there was very little difference. For greater than three subgroups the SR test was more powerful when reversal of treatment effect was present in very few (in particular one) subgroups but less powerful when the reversal was contained in several subgroups. (Note that power is a function, in this case, of δ_j / σ_j.) Therefore, from the perspective of drug development, where - apart from centre or country - subgroups tend to be few, the choice of test is to some extent unimportant, with perhaps the standardised range test being preferable in terms of its simplicity.

More recently, Pan and Wolfe (1997) published their simultaneous confidence interval approach - in effect an extension to the standardised range test. In simple terms, confidence intervals are constructed for each subgroup treatment difference and the null hypothesis is rejected if at least one confidence interval exists wholly greater than zero and at least one exists wholly less then zero. Each confidence interval is constructed using the following formula:

$$(L_j, U_j) = \left(D_j - z\left(\frac{1-P_E}{2}\right)\sigma_j, D_j + z\left(\frac{1-P_E}{2}\right)\sigma_j \right),$$

where $P_E = 2(1-\alpha)^{\frac{1}{J-1}} - 1$, is the confidence coefficient and z is the upper critical point of the standard normal distribution. In essence, as the number of subgroups increase, the width of the confidence interval increases to take this into account. Pan and Wolfe (1997) also extend their method to consider an *indifference* region where, in effect, some small degree of qualitative interaction (*d*) is allowed for in the null hypothesis. As a result, the modified null hypothesis is rejected if at least one confidence interval exists wholly greater than *d* and at least one wholly less then *-d*. Other variations on a theme include Yan (2004) and Wellek (1997). Yan builds upon the confidence interval approach of Pan and Wolfe by introducing a so-called tuning parameter that requires more than one pair of confidence intervals to fall wholly above *d* and below *-d* to reject the null hypothesis. Yan's approach is specifically directed towards treatment by centre interactions where the number of centres is large and a reversal of treatment effect is required in more than one centre to be clinically relevant. Wellek recognises the similarity of the problem with the issue of demonstrating equivalence and switches the hypotheses such that the alternative is now that no qualitative interaction exists. (Note that Wellek uses the extreme-value statistic approach as per the SR test and also introduces an indifference zone.)

Interestingly the concept of an indifference zone was noted originally in the Gail and Simon (1985) paper and was perhaps an early acknowledgement that the simple quantitative versus qualitative distinction does not completely address the clinical relevance of an interaction.

4.3 REGULATORY CONSIDERATIONS

Surprisingly, none of the recently issued regulatory guidance discusses the magnitude of the interaction effect despite comments directed towards inadequate power. Furthermore, there is no mention of methods to distinguish between quantitative and qualitative interactions and specifically no mention of formal hypothesis testing approaches such as those described by Gail and Simon (1985) and others in Section 4.2.

The Points to Consider on *Adjustment for baseline covariates* (CPMP/EWP/2863/99, 2003) essentially views the issue as being addressed in the earlier ICH E9 guideline. It simply highlights that in cases where an interaction is not, *a priori*, expected, *the primary analysis should only include the main effects for treatment and covariate.* In contrast, where an interaction is *expected then stratified randomisation and/or subgroup analyses should be pre-planned*, and the study should be powered to detect treatment differences within the separate subgroups. The investigation of interactions is considered to be very much an exploratory procedure although in an apparent change in emphasis states that the consistency of the treatment difference is an important consideration in the construction of *convincing evidence* of a *clinically useful effect.* The PtC notes that interaction tests often lack power and states that failure to reject the null hypothesis is not necessarily evidence of a clinically important interaction, while conversely, a statistically significant result should not be used in isolation to conclude clinical relevance. As such, its emphasis is that the evidence should be *examined carefully* and the primary analysis (excluding the

interaction) should be *interpreted cautiously* if the evidence points to an interaction. Indeed if the interaction is large or qualitative, the PtC deems that the interpretation of the primary analysis *may become impossible.*

The Points to Consider on *Multiplicity issues in clinical trials* (CPMP/EWP/908/99, 2002), addresses the specific issue of interactions to an even lesser extent, simply highlighting *that the evaluation of uniformity of treatment effects across subgroups is a general regulatory concern.* In addition it states that *a license may be restricted if unexplained strong heterogeneity is found in important sub-populations, or if heterogeneity of the treatment effect can reasonably be assumed but cannot be sufficiently evaluated for important subgroups.* So again, like the *Adjustment for baseline covariates* PtC, the message appears to be that the p-value resulting from a simple interaction test is almost superfluous due to concerns regarding power and that a more subjective approach is required.

In the context of treatment by centre interactions, ICH E9 points out that the inclusion of an interaction term in a model when the treatment effect is homogenous across strata is inefficient in terms of the evaluating the main effect of treatment. However when the effect is heterogenous across strata, ICH E9 describes the *interpretation of the main effect as controversial.* Specific advice is provided regarding treatment by centre interactions but actually very little in the general area of treatment by factor interactions. It states that where *interactions are anticipated or are of particular prior interest*, then the *planned confirmatory analysis* should include either subgroup analyses or the modelling of interactions. If interactions are not anticipated then such analyses are instead considered exploratory and involve the systematic addition of interaction terms to the primary model with complementary subgroup analyses. In these cases the results should be interpreted

cautiously and claims of a treatment effect based solely on this evidence *are unlikely to be accepted* by regulatory authorities.

In many respects it somewhat disappointing that the current guidance places so much emphasis on the perceived lack of power of interaction tests whilst ignoring the obvious point that power is irrelevant without consideration as to the magnitude of a clinically relevant interaction effect. Indeed the wishy-washy view that caution needs to be exercised almost regardless of the p-value, devalues any modelling approach to the problem. Surprisingly there is no mention of estimation and confidence intervals. An attempt at redressing this imbalance is presented in the next section when the clinical relevance of the interaction parameter is considered.

4.4 CLINICAL RELEVANCE OF THE INTERACTION PARAMETER

It is established good practice in clinical trial design that a primary study objective is specified and that the sample size is justified in terms of providing the power to detect a specific difference between treatments with respect to the primary endpoint whilst controlling the type I error. There are of course variations in approach depending on the design employed but the general principles of pre-specification and sample size justification apply. Determining the magnitude of the treatment difference often requires recourse to the scientific literature and in some cases regulatory guidelines. Although pragmatic solutions are frequently required the process - in all but new therapeutic areas - is mostly straightforward, if not a little subjective. Although determining a clinically relevant or plausible difference between treatments is certainly challenging it is however many times simpler than determining the difference between treatment differences that is clinical relevant.

As discussed earlier, the distinction between qualitative and quantitative forms of interactions does appear at first glance to have important treatment implications. For instance, if a treatment by factor interaction is present, and is quantitative in form, then the expectation when treating patients without regard to the level of this factor is that a patient will not be detrimentally impacted (with regard to the specific endpoint). That is, there is no expected loss. (Recall that the Gail and Simon (1985) orthants allow for zero treatment effects in one or more subgroups in the null hypothesis.) However if a qualitative interaction is present then the rational treatment of an individual patient must include consideration of the level of factor present.

As described by Gail and Simon (1985), when performing a statistical analysis of a RCT, instead of applying a pre-specified clinically relevant value for the interaction parameter it is more common simply to test for a statistically significant interaction and generate a p-value. (In fact it is extremely rare to find confidence intervals presented.) If the null hypothesis of no interaction is subsequently rejected then the logical next step is to test further in an attempt to rule out a qualitative interaction - although this could hardly be described as a universal approach in drug development as indicated by the absence of any mention in the regulatory guidance. However, as illustrated earlier, the traditional hypothesis testing approach ($H_0 : \Omega = 0$) may be uninformative, and a non significant result may not preclude the presence of a clinically relevant difference between the treatment differences. The obvious conclusion therefore is that current approaches to the evaluation of treatment by factor interactions are unsatisfactory.

Given the potentially important distinction between qualitative and quantitative interactions it would actually be useful if the interaction parameter itself could be used to facilitate the evaluation. Although the interaction parameter does not fully capture the

required distinction it is possible to set *post hoc* boundaries for the estimate that serve this purpose, as will be shown below.

Recall equation (4) that related the interaction parameter to the subgroup treatment difference and the overall weighted treatment difference:

$$\Omega = -2(\mu_{t2} - \mu_{r2}) + 2\Lambda .$$

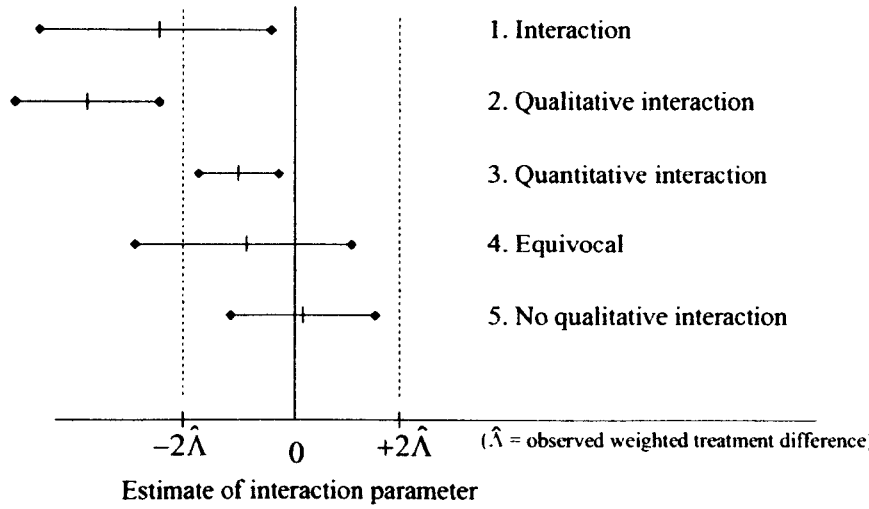
From this it can be determined that the interaction will be qualitative if $|\Omega| > 2|\Lambda|$, and that the corresponding boundaries for the transitions from quantitative to qualitative interaction are $(-2\Lambda,+2\Lambda)$ as illustrated in Table 4.II below. (Note also that these boundaries also hold when specified in terms of subgroup F=1 - that is, $\Omega = 2(\mu_{t1} - \mu_{r1}) - 2\Lambda$.)

Table 4.II: Relationship between the interaction parameter and weighted treatment difference

Factor		Λ	Ω	**Interpretation**
F=1	**F=2**			
2	2	2	0	No interaction
3	1	2	Λ=2	Quantitative
4	0	2	2Λ=4	Transition QNT/QL
5	-1	2	3Λ=6	Qualitative

Now, Figure 4.2 illustrates how estimated confidence limits for an interaction parameter – in association with the observed weighted treatment difference – could be used to interpret the estimate of the interaction.

Figure 4.2. Basic principles underlying confidence interval approach to the interpretation of the estimate of the interaction parameter



To declare that the interaction is statistically significant the only requirement is that the confidence interval (CI) excludes zero as shown in Case 1. In this instance, the range of values within the CI is consistent with both a quantitative and a qualitative interaction. In Case 2, the upper confidence limit is less than $-2\hat{\Lambda}$ and the interaction is purely qualitative. Case 3 represents a straightforward quantitative interaction where the presence of a qualitative interaction can be regarded as implausible. In Case 4, the range of the CI is consistent with all possibilities - no interaction, a quantitative interaction and a qualitative interaction. One might incorrectly assume that this situation always represents the case where the power of the test is low. However, if the treatments compared are on average similar, such that $2\hat{\Lambda}$ is small, then the CI could actually be relatively narrow - providing only a small range of plausible values for the interaction parameter. In Case 5, a qualitative interaction is ruled out as implausible, and the CI is consistent with either a quantitative interaction or no interaction. In this respect it is straightforward to see that a quantitative interaction effectively can only be ruled out in situations where the interaction

is clearly qualitative. (That is, the whole confidence interval is less than $-2\hat{\Lambda}$ or greater than $+2\hat{\Lambda}$.) Naturally this will correspond to cases where the estimate of the interaction parameter is simply large but will also include the scenario where $\hat{\Lambda} = 0$, in which case Ω must be either a zero or qualitative in nature.

As $|\hat{\Lambda}|$ increases, the result is that the range of values corresponding to a quantitative interaction also increases. This supports the view that qualitative interactions are much more plausible when treatments are expected to be, on average, equivalent. Furthermore, it is clear that as the overall treatment difference increases, then under the null hypothesis of no interaction, the chances of observing a statistically significant qualitative interaction is less likely - that is, a qualitative interaction distinguished from a quantitative one.

Now, consider the weighted treatment difference (Λ) again. At the design stage of a superiority trial, the specific treatment difference that the study aims to detect (δ) may be considered to be a good approximation for this parameter, and as such a qualitative interaction would now equate to the case were $|\Omega| > 2|\delta|$ leading to transition margins $(-2\delta, +2\delta)$. Now, in terms of the Wald statistic (maximum likelihood estimate/estimated standard error), it is interesting to note that a test of the overall treatment difference, with a Wald statistic of $\frac{\delta}{\sqrt{v}}$, would have identical power to a test for interaction when the interaction is qualitative, with a Wald statistic of $\frac{2\delta}{\sqrt{4v}}$. This observation would suggest that, *a priori*, superiority studies would tend to be adequately powered to detect interactions that are qualitative, so long as the study is adequately powered to detect the true overall treatment difference.

The generalisation of the approach to factors with more than two levels is not straightforward since it depends upon the number of levels, the effect size for each treatment by factor combination, the proportion of subjects categorised to each level and the type of effect tested for and present in the data (Kelly *et al*, 2005). (Kelly *et al* provide formulae to calculate the power and sample size for treatment by gene interactions based on the generalised linear model. They also adapt an alternative approach developed by Elston *et al* (1999) for binary responses to Gaussian outcomes.) For example, in terms of type of effect, Kelly *et al* illustrates a pattern of possible genotype response in terms of whether the different treatment-allele interaction effects are additive, dominant or recessive. In this case, a single diallelic locus with alleles A and a, provides three possible combinations: aa, aA and AA. An additive effect implies that there in an additive effect moving through the levels, a dominant effect implies an effect only with addition of A while a recessive effect implies an effect only with the AA combination.

Returning to the generalisation of the approach above, for illustrative purposes, if a factor has three levels with an equal proportion of subjects assigned to each combination, then the variance for the interaction for the simple comparison between two levels would be $6v$ since the variance of the treatment difference in each subgroup is now $3\upsilon^2 / 2n = 3v$. Similarly for four levels the variance of the interaction between a pair would be $8v$. Now δ remains a good approximation for Λ, where Λ now represents the weighted treatment difference across two factor levels, so the Wald statistic for a pairwise interaction becomes $\frac{2\delta}{\sqrt{6v}}$ and $\frac{\delta}{\sqrt{2v}}$ for three and four levels respectively. Thus to have the same power *a priori* as the test of the overall treatment difference, the magnitude of the pairwise interaction would need to be greater than δ by a factor of $\frac{\sqrt{3}}{\sqrt{2}}$ and $\sqrt{2}$ respectively.

Therefore, the result – that *a priori* a superiority study would tend to be adequately

powered to detect an interactions that was qualitative - cannot be generalised to interactions involving factors with more than two levels.

In the context of Figure 4.2, it is interesting to consider the views of Röhmel (1999) - a statistician at the Federal Institute for Drugs and Medical Devices in Germany - in relation to the investigation of treatment by centre interactions. Röhmel describes *the usually low power of test for quantitative interaction and the dramatically lower power to detect qualitative interaction.* However, as described above and as illustrated in Figure 4.2, it is important to make the distinction between detecting a qualitative interaction and distinguishing a qualitative interaction from a quantitative one. It is clear from Case 2, that if an interaction is qualitative (for illustration, this equates to $\hat{\Omega} < -2\hat{\Lambda}$), then compared to a smaller quantitative interaction (such as Case 3, where $-2\hat{\Lambda} > \hat{\Omega} > 0$), the power to detect a significant interaction (that is $\Omega \neq 0$) will actually be higher not lower. However it is indeed true, as is illustrated in Case 1, that the corresponding power to determine that the same interaction is qualitative and not quantitative would be much lower since the lower confidence limit would now need to be $< -2\hat{\Lambda}$ and not simply less than zero. (Oddly, Röhmel actually defines a qualitative interaction *as the existence of an overall superior therapy and the simultaneous existence of a subpopulation of patients in which the globally inferior therapy is actually better.* However Röhmel's definition is careless, since as was shown in Chapter Three, Simpson's paradox also describes the reversal of the overall effect in subgroups, and this does not represent a qualitative interaction. Röhmel's definition is only valid therefore if the claim of overall superiority that he refers to is based upon a stratified analysis that includes the corresponding subgroup defining factor.)

As has been illustrated, the overall weighted mean treatment difference (Λ) can be viewed as providing a convenient way of classifying the interaction parameter as either quantitative or qualitative. However since Λ cannot be determined *a priori* and must be estimated from the data, the boundaries should more accurately be described as "pseudo margins" as they are by their nature data dependent. An alternative and preferable approach would be the *a priori* specification of margins using a non-data dependent method of elicitation. Now, the concept of using pre-specified margins to evaluate treatment differences is well established in the area of bioequivalence. In this respect regulatory authorities provide specific values for these margins when pharmaceutical companies design studies to compare a generic drug with an original approved drug which is off patent. The method of evaluation has also expanded into the clinical environment when assessing therapeutic equivalence - the subject of Chapter Five. In therapeutic equivalence, two margins are established, $(-m_1,+m_2)$ say, which define a range of values for the primary endpoint within which the difference between two active treatments is considered clinically equivalent. Although there are no widely established values for these margins in equivalence methodology, a CPMP concept paper (CPMP/EWP/2158/99) has suggested a margin *of one half to one third of the established superiority of the comparator to placebo*. Similarly Phillips *et al* (2000) have proposed a margin *less than half the difference between active and placebo*. In truth this area of margin specification remains controversial (as will be demonstrated in Chapter Five) and the most recent draft guidance (Points to consider on the choice of non-inferiority margin, (CPMP/EWP/2158/99 draft)) is much less prescriptive than had been earlier indicated in the concept document. However one could borrow, or build upon, the originally suggested margins emanating from the area of therapeutic equivalence for the difference between active treatments, and apply these to the difference between the treatment differences.

The most straightforward application would be for a placebo controlled superiority trial. In this case, the study would have pre-specified a treatment difference at the design stage for the primary endpoint, δ, and so margins for the interaction could be simply defined as $(-\delta/2,+\delta/2)$ or $(-\delta/3,+\delta/3)$. However the earlier observation that a superiority trial could have *a priori* adequate power to detect a qualitative interaction might also lend support for margins much wider than those proposed above - that is, $(-2\delta,+2\delta)$. Taking this one step further, when describing their indifference region (see Section 4.4), Pan and Wolfe (1997) assign a value of 1% to d, for example, as a means of including some small degree of qualitative interaction in the null hypothesis when comparing response percentages. This would point to margins of $(-[2\delta+1],+[2\delta+1])$ for data analysed on the percentage scale. However according to Senn (2003), the threshold for a clinically relevant interaction *clearly cannot be larger* than the clinically relevant difference overall, δ, and *ought to be less*. This would suggest margins no wider than $(-\delta,+\delta)$ and one can see from Table 4.II that, if δ is replaced by Λ, then this equates to the size of effect in one subgroup being three times larger than the effect in the complement subgroup – a difference that intuitively seems worthy of note. For an active controlled equivalence study, appropriate symmetric margins $(-m,+m)$ would have already been pre-specified and these margins could simply be applied to the interaction also.

Rather like therapeutic equivalence and non-inferiority, perhaps it is impossible to be too prescriptive in terms of specifying the magnitude of the margins and instead these should be determined on the basis of both the specific hypothesis being tested (including the nature of the reference treatment) and the specific therapeutic areas. However it is clear that the subject would benefit from considered thought prior to study initiation.

If margins could be established for the interaction parameter, the next step would be to consider how to evaluate the resulting estimate against these margins. The approach adopted in bio- and therapeutic equivalence is to regard the confidence interval for the estimate as representing a range of plausible values for the parameter. It follows that if the confidence interval is wholly contained within the range of the margins then it is considered reasonable to conclude that the interaction parameter is no greater than the margins and that any difference between the treatment differences is clinically irrelevant. However this approach is directed at evaluating primary parameters for which the study has been adequately powered whereas the evaluation of treatment by factor interactions is usually of a secondary or exploratory nature. In this respect such an approach may simply produce an equivocal result and show that a clinically relevant difference cannot be ruled out.

An alternative approach might be to adopt Bayesian type thinking which could be applied to produce a more useful way of tackling the problem of estimate imprecision. Indeed in a more general sense, Peto (1982) stresses the importance of using prior information when interpreting interactions, and so a Bayesian approach also provides an opportunity to incorporate informative prior information to aid the interpretation of the data. This Bayesian approach is discussed in the next section.

4.5 A BAYESIAN APPROACH

4.5.1 General approach

One potential approach would be to calculate the posterior probability that the interaction parameter lies within the interval defined by the margins, $(-\varepsilon,+\varepsilon)$ say. That is, $P(-\varepsilon \leq \Omega \leq +\varepsilon)$. Using a Bayesian approach the interaction parameter, Ω, is regarded as random, and this can be contrasted with the classical approach whereby the data are

regarded as random but the parameter, Ω, is fixed but unknown (Lee, 1989). Furthermore, prior beliefs regarding Ω are modified once data are observed to arrive at a posterior probability distribution using the formulation that the posterior distribution is proportional to the prior distribution multiplied by the likelihood (that is, the data).

Bayesian approaches for continuous and binary outcome will now be developed in the next two sub-sections for the simple case of two treatments with a two-level single factor.

4.5.2 Continuous outcomes

Consider the linear model

$$x_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

where $\mu_{ij} = \alpha + \tau_i + \phi_j + (\tau\phi)_{ij}$ and subject (k) is subject to the effects of treatment τ_i (i=t,r) and factor ϕ_j (j=1,2) and that there is also a background level, represented by the constant term α. Furthermore the errors ε_{ijk} are independent and identically distributed with constant variance.

Now, the interaction between treatment and factor is the difference between the treatment difference for factor level F=1 (μ_{t1} - μ_{r1}) and the treatment difference for factor level F=2 (μ_{t2} - μ_{r2}), and so can be formally specified as:

$$(\mu_{t1} - \mu_{r1}) - (\mu_{t2} - \mu_{r2}),$$

or equivalently,

$$\mu_{t1} - \mu_{r1} - \mu_{t2} + \mu_{r2}$$

Furthermore, since it follows from general theory that any contrast of the means $\sum_i \sum_j c_{ij}\mu_{ij}$ with $\sum_i \sum_j c_{ij} = 0$ can be estimated by $\sum_i \sum_j c_{ij}\bar{x}_{ij}$, the estimate of the interaction is simply $(\bar{x}_{t1} - \bar{x}_{r1} - \bar{x}_{t2} + \bar{x}_{r2})$.

Also from general theory, the variance of the contrast is $\sigma^2 \sum_i \sum_j c_{ij}^2 n_{ij}^{-1}$. Now, if it is assumed that σ^2 is unknown, but equal for all four treatment by factor combinations, then it can be estimated as the weighted mean of the individual s_{ij}^2 as follows:

$$s^2 = \frac{(n_{t1}-1)s_{t1}^2 + (n_{t2}-1)s_{t2}^2 + (n_{r1}-1)s_{r1}^2 + (n_{r2}-1)s_{r2}^2}{(n_{t1}-1)+(n_{t2}-1)+(n_{r1}-1)+(n_{r2}-1)},$$

where s_{ij}^2 is calculated as $\frac{\sum(x_{ijk}-\bar{x}_{ij})^2}{n_{ij}-1}$

This can then be re-arranged to give,

$$s^2 = \frac{\sum(x_{t1k}-\bar{x}_{t1})^2 + \sum(x_{t2k}-\bar{x}_{t2})^2 + \sum(x_{r1k}-\bar{x}_{r1})^2 + \sum(x_{r2k}-\bar{x}_{r2})^2}{n_{t1}+n_{t2}+n_{r1}+n_{r2}-4}$$

It then follows from the variance of the contrast ($\sigma^2 \sum_i \sum_j c_{ij}^2 n_{ij}^{-1}$) that the standard error of the interaction is

$$SE(\hat{\Omega}) = \sqrt{s^2\left(\frac{1}{n_{t1}}+\frac{1}{n_{t2}}+\frac{1}{n_{r1}}+\frac{1}{n_{r2}}\right)}$$

which follows a t-distribution with $n_{t1} + n_{t2} + n_{r1} + n_{r2} - 4$ degrees of freedom

For each treatment by factor combination, non-trivial prior information can be expressed for the unknown parameter μ in terms of the parameters η and ς from a normal distribution, that is,

$$\mu \sim N(\eta, \varsigma^2).$$

(Note to aid clarity the subscripts (i and j) have been dropped.) Now, if n observations from a normal distribution have been generated then,

$$x \sim N(\mu, \sigma^2)$$

but with the restriction that the variance, σ^2, is known and,

$$\bar{x} \sim N(\mu, \sigma^2/n)$$

which leads to the posterior distribution (Lee, 1989),

$$\mu \sim N(\theta,\ \xi^2),$$

where,

$$\xi^2 = \left\{\varsigma^{-2} + \left(\sigma^2 / n\right)^{-1}\right\}^{-1}$$

$$\theta = \xi^2\left\{\eta/\varsigma^2 + \bar{x}/\left(\sigma^2/n\right)\right\}.$$

These can then be re-arranged to give:

$$\xi^2 = \frac{1}{\varsigma^{-2} + \left(\sigma^2/n\right)^{-1}}$$

$$\theta = \frac{\eta\varsigma^{-2}}{\varsigma^{-2} + \left(\sigma^2/n\right)^{-1}} + \frac{\bar{x}\left(\sigma^2/n\right)^{-1}}{\varsigma^{-2} + \left(\sigma^2/n\right)^{-1}}.$$

In these forms, the posterior precision (defined as the inverse of the variance) can be described as being the sum of the prior precision and the data precision while the posterior mean can be described as a weighted mean of the prior mean and observed mean.

Note that for the observed data, this method assumes that the variance is known so in practice σ^2 is substituted by the estimate s^2. For scenarios relating to the evaluation of a treatment by factor interaction it is likely that the sample sizes will be reasonably large – typically resulting from large phase III confirmatory studies – so this method will be valid and differ little from a solution involving an unknown variance assumption.

Accordingly, interest centres upon the distribution of the interaction but this is now a posterior distribution which, using general theory, is normally distributed as follows:

$$\mu_{t1} - \mu_{r1} - \mu_{t2} + \mu_{r2} \sim N\left(\theta_{t1} - \theta_{r1} - \theta_{t2} + \theta_{r2}, \xi_{t1}^2 + \xi_{r1}^2 + \xi_{t2}^2 + \xi_{r2}^2\right)$$

Now, Simon and Freedman (1997) applied Bayesian methods to the analysis of the two treatment factorial design and in particular to the evaluation of the interaction between the

two treatments. The evaluation of interactions is a problem that the factorial model shares with the treatment by factor model although there are also a number of differences. In the two treatment factorial model, subjects are randomised to one of four treatment combinations representing the presence or absence of each of two treatments, and assumptions regarding the interaction parameter are integral to study design. In this context, the factorial model studies the joint effect of two randomised treatments and assumptions regarding the presence or absence of interaction have important implications for the sample size. For instance, if no interaction is assumed then the factorial model provides an efficient design - that is, two treatments can be studied jointly using fewer subjects than if each were studied separately. However if it is important to evaluate the effect of giving two treatments together then such a trial will be under powered to detect an interaction, unless the sample size is increased accordingly. Furthermore treatment interactions of both a positive and negative nature are possible. For instance two treatments may combine to produce a synergistic effect although it is quite plausible that the combined effect may be no greater than the individual effect of each treatment alone.

In contrast the evaluation of treatment by factor interactions is more exploratory in nature and although a formal investigation of the interaction between treatment and a stratification factor may be planned at the design stage, many other factors might be evaluated in the same study. Another important distinction is that subjects are not randomised to a specific level of a factor – rather factors are either inherent characteristics of the subject (such as gender) or background features (such as disease severity). This has important implications since there is no guarantee that male subjects, say, have been sampled in the same way as females – perhaps due to protocol exclusions relating to child bearing potential status. Finally the concept of positive (or synergistic) and negative

interaction effects is not relevant to treatment by factor interactions since unlike factorial designs, different levels of a factor will usually have equal importance.

Nevertheless, the factorial model proposed by Simon and Freedman (1997) is broadly applicable to the evaluation of treatment by factor interactions and provides an alternative parameterisation that has some advantages over the basic approach described earlier. Consider a model with two treatments (represented by different levels of τ coded –1 or 1) and a two-level factor (represented by different levels of ϕ also coded –1 or 1) that takes the form:

$$x = \beta_0 + \beta_1\tau + \beta_2\phi + \beta_3(\tau\phi) + \varepsilon,$$

where the error terms ε are N(0, σ^2) and independently distributed. Now, with this parameterisation, the main effect of treatment is given by $2\beta_1$, the main effect of the factor by $2\beta_2$, and the treatment by factor interaction by $4\beta_3$ and it follows that maximum likelihood estimates of the parameters are given by the four contrasts:

$$\hat{\beta}_0 = (\bar{x}_{t1} + \bar{x}_{r1} + \bar{x}_{t2} + \bar{x}_{r2})/4$$

$$\hat{\beta}_1 = (\bar{x}_{t1} - \bar{x}_{r1} + \bar{x}_{t2} - \bar{x}_{r2})/4$$

$$\hat{\beta}_2 = (\bar{x}_{t1} + \bar{x}_{r1} - \bar{x}_{t2} - \bar{x}_{r2})/4$$

$$\hat{\beta}_3 = (\bar{x}_{t1} - \bar{x}_{r1} - \bar{x}_{t2} + \bar{x}_{r2})/4,$$

each with variance $\sigma^2/4n$.

Simon and Freedman assume a prior distribution for the vector of parameters, $\underline{\beta}$, of the form $N(\underline{\mu}, \Sigma)$ and, in accordance with Lindley and Smith (1972), show that if

$\underline{\hat{\beta}} \mid \underline{\beta} \sim N(\underline{\beta}, C)$, then the resulting posterior distribution is $\underline{\beta} \mid \underline{\hat{\beta}} \sim N(\mathrm{B}\underline{b}, \mathrm{B})$, where $\mathrm{B}^{-1} = C^{-1} + \Sigma^{-1}$ and $\underline{b} = C^{-1}\underline{\hat{\beta}} + \Sigma^{-1}\underline{\mu}$. $C = (\sigma^2 / 4n)\mathrm{I}$, where I represents the identity matrix, and assuming that the prior distributions for all four parameters are independent then $\Sigma = diag(\upsilon_0^2, \upsilon_1^2, \upsilon_2^2, \upsilon_3^2)$.

As such the posterior distribution for the interaction parameter (β_3) is normally distributed with mean: $4\dfrac{(4n/\sigma^2)\hat{\beta}_3 + (\mu_3/\upsilon_3^2)}{(4n/\sigma^2) + (1/\upsilon_3^2)}$ and variance $\dfrac{16}{(4n/\sigma^2) + (1/\upsilon_3^2)}$

4.5.3 Binary outcomes

Continuing with the simple case of two treatment groups and one baseline factor with two levels, then for some binary outcome the true proportion (proportion of responders, say) for each treatment by factor combination is represented by π_{ij} with corresponding estimate x_{ij}/n_{ij}.

		Factor	
		F= 1	F= 2
Treatment	T= t	x_{t1}/n_{t1}	x_{t2}/n_{t2}
	T= r	x_{r1}/n_{r1}	x_{r2}/n_{r2}

Corresponding prior probabilities of response can be expressed for each treatment by factor combination in terms of the parameters, γ and φ, of a beta distribution. That is, $\pi_{ij} \sim Be(\gamma_{ij}, \varphi_{ij})$.

		Factor	
		F= 1	F= 2
Treatment	T= t	$\gamma_{t1}/(\gamma_{t1} + \varphi_{t1})$	$\gamma_{t2}/(\gamma_{t2} + \varphi_{t2})$
	T= r	$\gamma_{r1}/(\gamma_{r1} + \varphi_{r1})$	$\gamma_{r2}/(\gamma_{r2} + \varphi_{r2})$

As such, each prior has the form $p(\pi) \propto \pi^{\gamma-1}(1-\pi)^{\varphi-1}$ where $(0 \leq \pi \leq 1)$. Now, the beta distribution is conjugate for the binomial likelihood (Lee, 1989) so that the posterior also has a beta distribution of the form

$$p(\pi / x) \sim \pi^{\gamma+x-1}(1-\pi)^{\varphi+n-x-1}$$

That is, if the prior distribution is beta, i.e. $\pi_{ij} \sim Be(\gamma_{ij}, \varphi_{ij})$ then the posterior distribution with also follow a beta distribution, i.e. $\pi_{ij} \sim Be(\alpha_{ij}, \beta_{ij})$, where $\alpha_{ij} = (\gamma_{ij} + x_{ij})$ and $\beta_{ij} = (\varphi_{ij} + n_{ij} - x_{ij})$. (Conjugate refers to the general case where a family of prior distributions leads to the same class of posterior distributions.)

4.5.3.1 Odds ratio

For each treatment by factor combination the odds is defined as:

$$\lambda_{ij} = \pi_{ij} / (1 - \pi_{ij})$$

and if $\pi_{ij} \sim Be(\alpha_{ij}, \beta_{ij})$ then the log odds, $\log(\lambda_{ij})$, is close to having Fisher's z distribution (Lee, 1989). That is,

$$\frac{1}{2}\log\lambda + \frac{1}{2}\log\left(\frac{\beta}{\alpha}\right) \sim z_{2\alpha.2\beta}$$

From the properties of the z distribution it follows that,

$$E(\log\lambda) \cong \log\{(\alpha - \tfrac{1}{2})/(\beta - \tfrac{1}{2})\}$$

$$Var(\log\lambda) \cong \alpha^{-1} + \beta^{-1}$$

Hence the log odds ratio, $\log(\lambda_{t1} / \lambda_{r1}) = \log(\lambda_{t1}) - \log(\lambda_{r1})$, for factor F=1 also has an approximate Normal distribution

$$N\left(\log\{(\alpha_{t1} - \tfrac{1}{2})(\beta_{r1} - \tfrac{1}{2})/(\beta_{t1} - \tfrac{1}{2})(\alpha_{r1} - \tfrac{1}{2})\}, \alpha_{t1}^{-1} + \beta_{t1}^{-1} + \alpha_{r1}^{-1} + \beta_{r1}^{-1}\right)$$

or more approximately

$$N\left(\log\{\alpha_{t1}\beta_{r1} / \beta_{t1}\alpha_{r1}\}, \alpha_{t1}^{-1} + \beta_{t1}^{-1} + \alpha_{r1}^{-1} + \beta_{r1}^{-1}\right)$$

The approximation is likely to be reasonable if all entries in the 2x2 table are at least 5 (Lee, 1989). Now, to evaluate a treatment by factor interaction one simply needs to compare the odds ratio from subgroup F=1 with the respective odds ratio from F=2. The most straightforward approach is to simply compare the difference in the log odds ratios. That is,

$$\log(\lambda_{t1}/\lambda_{r1}) - \log(\lambda_{t2}/\lambda_{r2}) = \{\log(\lambda_{t1}) - \log(\lambda_{r1})\} - \{\log(\lambda_{t2}) - \log(\lambda_{r2})\},$$

and this also has an approximately Normal distribution

$$N\left(\log\{\alpha_{t1}\beta_{r1}\beta_{t2}\alpha_{r2}/\beta_{t1}\alpha_{r1}\alpha_{t2}\beta_{r2}\}, \{\alpha_{t1}^{-1} + \beta_{t1}^{-1} + \alpha_{r1}^{-1} + \beta_{r1}^{-1} + \alpha_{t2}^{-1} + \beta_{t2}^{-1} + \alpha_{r2}^{-1} + \beta_{r2}^{-1}\}\right)$$

4.5.3.2 Difference in proportions

The mean and variance of a beta distribution are given by

$$E(\pi) = \alpha/(\alpha+\beta)$$

$$Var(\pi) = \alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$$

Now, if both α and β are at least 10 then the beta distribution can be approximated by a Normal distribution with the same mean and variance (Lee, 1989). It follows, therefore, that $\pi_{t1} - \pi_{r1}$ has a Normal distribution with mean $\alpha_{t1}/(\alpha_{t1}+\beta_{t1}) - \alpha_{r1}/(\alpha_{r1}+\beta_{r1})$ and variance $[\alpha_{t1}\beta_{t1}/(\alpha_{t1}+\beta_{t1})^2(\alpha_{t1}+\beta_{t1}+1)] + [\alpha_{r1}\beta_{r1}/(\alpha_{r1}+\beta_{r1})^2(\alpha_{r1}+\beta_{r1}+1)]$. Again, it is then straightforward to extend this to the distribution for the difference between the subgroup treatment differences (i.e. treatment by factor interaction), assuming that all the α_{ij} and β_{ij} terms are at least 10.

4.5.4 Incorporating prior information

Spiegelhalter *et al* (1994) discuss a range of prior distributions that can be applied to a clinical trial setting and in fact explicitly encourage a full range of prior distributions to be applied to data set as opposed to the assignment of a single prior. Spiegelhalter *et al* view

this family of priors as representing a range of perspectives based upon evidence that is external to the current clinical trial and categorise the priors as reference, clinical, sceptical and enthusiastic.

4.5.4.1 Reference priors

The essential aim of a reference prior is to offer the least information possible such that the likelihood is essentially untouched. In this respect a reference prior is the least subjective of all priors but is to a certain extent unrealistic as a result. Spiegelhalter *et al* (1994) describe the reference prior as one that provides a baseline from which the impact of other priors can be judged. Lee (1989) describes a reference prior as *neutral* - that is, *the views of someone who had no strong beliefs a priori.* (A pragmatic alternative with the same aim is simply to interpret the likelihood in a Bayesian way as suggested by practical Bayesian statisticians (for instance, Professor Peter Freeman, personal communication).)

Although simple in concept, the choice of reference priors is not necessarily straightforward as illustrated by the application to binary outcome data. A uniform distribution in which no value is more likely than another is intuitively appealing – however in many cases the density function would not integrate to 1 and would be described as improper. (A proper density by contrast would integrate to 1.) In effect there are three main choices when selecting a conjugate reference prior for the binomial likelihood – although others have been suggested. These are the beta priors: Haldane's prior, Be(0,0), which is an improper density but is a uniform prior for the log odds; the arc-sine prior, Be(½, ½), that is a proper density and is a uniform prior for $\sin^{-1}\sqrt{\pi}$; and Bayes' postulate, Be(1,1), that is also a proper density. Lee (1989 provides a full

discussion of the properties of these three priors although in practice unless the number of observed data is very few, the choice is irrelevant.

4.5.4.2 Informative priors: clinical, sceptical and enthusiastic.

Using an informative prior is essentially equivalent to simply adding data from outside of the trial to the data observed in the trial – and weighting the two sets of data appropriately. Spiegelhalter *et al* (1994) describe a clinical prior as one that is either based on the opinion of one or more well informed individuals or is derived from historical data – perhaps following a meta analysis. In considering the former, Senn (in the discussion of the Spiegelhalter *et al* (1994) paper) asks: *Is this 'group clinical opinion' not an example of information which is worse than useless?* Machin, in the same discussion, suggests that *at best it will be a compound of anecdotal information, the published literature and personal experience* and generally views such priors as representing optimistic belief. Similarly Evans – who would subsequently join the MCA - comments further that *the beliefs that the clinicians have are very often entirely unsupported* and that *it is important to distinguish between belief and belief supported by evidence.* Indeed in a regulatory environment where confirmatory evidence of effect is being presented it is hard to imagine the acceptability of such opinionated priors. Priors based on previous data appear less controversial – Evans stating *that meta-analysis priors should be a requirement for part of the introduction to papers.* Lewis – who would also subsequently join the MCA - adds: *I want their [the investigators] views to reflect previous data.* The use of historical data is not necessarily straightforward however and it is important that these data are not simply viewed as being exchangeable (that is, of equal weight) with the data generated prospectively from the study itself. In this respect it is usual to shrink the treatment effect and/or increase the variance of the distribution to allow for the potential heterogeneity of these data in comparison with the prospective data. Specific sources of prior information could also include the use of pharmacokinetic data - identifying differences in drug

distribution between subgroups for specific drugs - and minimum inhibitory concentrations – identifying which drugs had the greatest *in vitro* potential against different bacteria in anti-infective trials. In both of these cases such information would be helpful in establishing clinical priors for evaluating a treatment by subgroup interaction.

The sceptical prior is intended to represent belief that a large treatment difference is unlikely. Typically the prior distribution would be symmetric around a treatment difference of zero and would have a variance that represented a range of plausible treatment differences. Spiegelhalter *et al* (1994) suggest *a prior equivalent to already having observed a quarter of the trial with a zero treatment difference* and use the treatment difference specified for the alternative hypothesis as the basis for the calculation. The impact if using such a prior would be to dampen any observed treatment difference. However as pointed out by Stephen Senn (personal communication), the curious feature of this approach is that an increase in the planned enrolment leads to more precise prior information. In this respect the larger the trial, the greater the prior information.

In contrast the enthusiastic prior is aimed at offsetting the impact of the sceptical prior. In this case the distribution would still be symmetric (with the same variance as the sceptical prior) but would now be centred on the treatment difference for the alternative hypothesis. Although many other variations are possible, Spiegelhalter *et al* were clearly aiming at establishing the concept of standard or 'off the shelf' priors that would become widely accepted.

In the context of treatment by subgroup interactions, the standard definitions to not transfer readily however. Thinking in terms of a prior for the interaction parameter itself (as per the model of Simon and Freeman (1997) described in Section 4.5.2), whether a

prior centred on zero would represent the sceptical view would depend on whether one was sceptical that an intreaction was present or absent. Certainly a prior centred on zero would dampen the possibility of finding a large interaction when considering the posterior and could be appropriate for exploratory analyses such that the possibility of false positive findings would be reduced. However from a regulatory perspective such an approach may not represent scepticism. In contrast an enthusiastic prior would point to the presence of an interaction in a specific direction (otherwise some sort of bi-modal distribution would be required that was enthusiastic in terms of a difference (interaction) in either direction but sceptical at the point of no interaction) and would most likely be based on historical data – pharmacokinetic differences between subgroups, for instance. It would also require an alternative hypothesis to be specified if the Spiegelhalter *et al* approach were followed. However such alternative hypothesis are seldom specified in terms of a specific magnitude of effect.

4.5.4.3 MCMC methods and non conjugate priors.

The introduction of Markov chain Monte Carlo (MCMC) methods, in particular the Gibbs sampler (Geman and Geman (1984) and Gelfand and Smith (1990)), has meant that many previously intractable Bayesian problems have become much more straightforward to solve – specifically in comparison with algebraic or numerical integration methods. Such iterative simulation of posterior distributions provides greater flexibility and practical implementation of the Gibbs sampler has been greatly enhanced by the introduction of BUGS (Bayesian inference Using Gibbs Sampling) computer software - including WinBUGS, Version 1.4 for the personal computer (Spiegelhalter *et al*, 2001). One particular feature of Gibbs sampling is that the technique does not restrict the selection of priors to the conjugate family. For instance in Section 4.5.3, the prior distribution for the binomial likelihood would not be restricted to the beta distribution. Furthermore BUGS

software is very flexible – affording the opportunity to specify a range of model formulations and parameterizations. Indeed it may be used simply to undertake straightforward maximum likelihood analysis. Smith and Roberts (1993) and Gelman and Rubin (1996) provide overviews of MCMC methods.

4.5.5 Further regulatory considerations in relation to Bayesian methods

Both ICH E9 and the earlier CPMP *Note for Guidance* (III/3630/92-EN) entitled *Biostatistical Methodology in Clinical Trials in Applications for Marketing Authorisations for Medicinal Purposes* (Lewis *et al*, 1995), on which ICH E9 was based, refer to Bayesian methods in their introductory sections but no specific mention is made thereafter. The CPMP *Note for Guidance* states:

> *Although this Note for Guidance is written largely from a classical (frequentist) viewpoint, the use of Bayesian or other well-argued approaches is quite acceptable.*

Meanwhile ICH E9 states:

> *Because the predominant approaches to the design and analysis of clinical trials have been based on frequentist statistical methods, the guidance largely refer to the use of frequentist methods when discussing hypothesis testing and/or confidence intervals. This should not be taken to imply that other approaches are not appropriate: the use of Bayesian and other approaches may be considered when the reasons for their use are clear and when the resulting conclusions are sufficiently robust.*

Grieve (in the discussion of Senn (2000)) has stated that *there is a feeling that regulatory authorities look less favourably on Bayesian approaches* – and concludes that ICH E9's

endorsement of Bayesianism is lukewarm. Senn (2000) describes *regulatory nervousness at the thought of officially making use of subjective inputs.* Earlier Grieve (in the discussion of the Spiegelhalter *et al* (1994) paper) identified *an explicit acceptance of Bayesian Methodology* in early phase drug development (that is, in the area of pharmacokinetics and pharmacodynamics) and indeed had himself been joint author of an influential paper in the mid 1980s describing the use of Bayesian methods in the pharmaceutical industry as applied to pre-clinical and early phase clinical trials (Racine et al, 1986). Indeed there is increasing interest in the use of Bayesian methods within the pharmaceutical industry in this area – not so much in terms of providing convincing evidence for regulatory approval of experimental treatments but rather directed at the internal decision making framework (dose selection, for instance) prior to performing the large confirmatory studies required for registration. Spiegelhalter *et al* (1994) alludes to this point: *Thus it appears completely sensible that a drug company can have its own prior evidence and loss function for its internal decision-making.*

Spiegelhalter *et al* (1994) noted that since the pharmaceutical industry follows the regulatory lead, then the use of Bayesian methods in this context would only happen if the regulatory agencies actually provided some encouragement. They further suggest that the sceptical prior would be the choice of the regulators in their role of *public watch-dog* where a degree of scepticism is consistent with such a role. It is not clear however that this approach would be consistent with the desire for direct encouragement since the impact might be to raise the hurdle in terms of what constitutes convincing evidence of an effect in confirmatory trials. Lewis (in the discussion of Senn (2000)) states that pharmaceutical industry *want their technology to be non-controversial and reliable so that drug development plans can be executed cleanly and efficiently.*

In the discussion of the Spiegelhalter *et al* (1994) paper in the mid 1990s, a number of eminent statisticians with regulatory connections commented. Lewis was keen to encourage the separation of results (based on data) from the interpretation (where there was potential for Bayesian thinking). Similarly Ellenberg (a statistician at the FDA) focussed on the role that Bayesian methods could have *in improving and enhancing the interpretability of trial results* – although she was doubtful whether they would replace current methods. She described the sceptical prior concept as intuitively appealing whilst highlighting that its arbitrariness that would lead to *intensive discussion* with Sponsors in terms of the *size of the 'handicap'*. This view was supported by Gail who highlighted that it was important for conventions to be widely accepted if they were subsequently to have widespread use. Ashby perceived *an increasing willingness among statisticians involved in drug regulation to explore the use of Bayesian approaches* – stressing a desire that when the statistics guidelines were updated that they provided more explicit advice in this area. To date this has not been the case since none of the European statistically related guidelines (mostly points to consider documents) issued since 2000 (as detailed in Chapter 1, Section 1.3.2) include any reference to Bayesian methods. This may partly be due to the fact that these documents (PtC) are aimed at addressing issues that have been identified in recent submissions and if submissions have not included Bayesian methods then it follows that such methods would not be the subject of clarification. Lewis and Facey (1998) and Lewis *et al* (2001) describe statistical shortcomings and submissions and the impact of ICH E9 respectively – yet neither includes a reference to Bayesian methods.

However the most recent CHMP draft guideline that includes substantial statistical content - the *Guideline on Clinical Trials in Small Populations* (CHMP/EWP/83561/2005, 2005) – perhaps represents a paradigm shift. This draft guideline is the first document to encourage actively the application of Bayesian methods in relation to the submission of

evidence for consideration of drug approval – with the caveat that Sponsors should seek scientific advice from the CHMP at the design stage. At the heart of the guideline is a clear acknowledgement that in rare diseases, where few subjects are available for clinical research, *it is imperative that the most efficient and informative analytical methods should be used.* It highlights that the use or more complex methods (including Bayesian modelling) usually involve extra assumptions which may not be verifiable, and as such stresses the importance of sensitivity analyses to assess the robustness of the conclusions. The complete sub-section on Bayesian methods states:

> *Bayesian methods are a further source of 'adding assumptions' to data. They are a way to formerly [sic] combine knowledge from previous data or prior 'beliefs' with data from a study. Introducing prior beliefs is often a concern in drug regulation. However, being able to use knowledge of likely effects of drugs due to their chemical form, likeness to other existing compounds, mechanism of action, and so on, is a very valuable addition to sparse data. As with sensitivity analyses mentioned above, a variety of reasonable prior distributions should be used to combine with data from small studies to ensure that conclusions are, at least, reasonably data-dependent and not almost entirely belief-dependent.*

Of note is the explicit reference above to the apprehension that the regulatory agencies have when combining prior beliefs with the prospectively collected data from a clinical trial. The draft guideline also presents limited details of a published example of Bayesian methods in practice in an Appendix. Of interest here is that a sceptical prior is defined as assuming that the test treatment is worse than the reference, a neutral prior assumes the test treatment has no effect at all and the enthusiastic prior assumes the test treatment has a *predefined realistic effect*. Although different to the definitions used earlier, the same principle of using a range of priors to perform a sensitivity analysis to assess robustness is evident.

Grieve (in the discussion of Senn (2000)) is *now more optimistic that Bayesian methods will gain a greater foothold then hitherto* and describes a future where data driven priors (and utility functions) will be detailed in study protocols with full justification.

To illustrate how some of the concepts that have been raised in the previous sections of this Chapter could be applied in practice, some published data where an interaction was detected have been selected. The re-analysis of these data is the subject of the next section.

4.6 AN EXAMPLE

Swarbrick *et al* (1996) published the results of a comparison of lansoprazole 30mg once daily with ranitidine 300mg twice daily in the treatment of oesophageal stricture where the endpoint was the proportion of subjects requiring re-dilatation during the 12-month treatment period. The sample size was chosen on the basis of detecting a reduction in the re-dilatation proportion of 25 percentage points with lansoprazole compared with ranitidine, which had an expected re-dilatation percentage of 50%.

The primary analyses followed the ITT principle and included all 158 randomised subjects and the following results were reported. The observed re-dilation percentages were 43.8% (35/80) for ranitidine and 30.8% (24/78) for lansoprazole, although the difference of 13.0% (95% CI: -3.2, 29.2) was not statistically significant (χ^2 test, p=0.092). (Note that the observed difference was markedly smaller than that specified in the alternative hypothesis.) A total of eight prognostic factors (including centre) were examined and one statistically significant interaction was reported - that between treatment and baseline drinking status (Breslow-Day homogeneity test, p=0.017) as shown in Table 4.III.

Table 4.III: Re-dilation proportions by treatment group and drinking status

	Drinker	Non-drinker
Lansoprazole	25.6% (11/43)	37.1% (13/35)
Ranitidine	54.2% (26/48)	28.1% (9/32)

The authors had no immediate explanation for what appeared to be a qualitative type interaction and stated that it might simply be a chance finding. Of note is that the primary presentation of the re-dilation percentages is given on the difference scale (the odds ratio is also presented but without confidence limits), yet the interaction is investigated on the odds scale using the Breslow-Day homogeneity test. The interaction would appear qualitative on either scale, however. Swarbrick *et al* (1996) did not investigate the apparent interaction further so it is interesting to re-analyse these data using some of the methods described earlier in this chapter.

Table 4.IV: Re-analysis of Swarbrick *et al* (1996)

Parameter estimated	OR	LogOR	SE (LogOR)	Wald 95% CI (LogOR)	p-value
Unadjusted Treatment	1.75	0.560	0.333	(-0.093, 1.213)	0.093
Adjusted Treatment	1.72	0.547	0.334	(-0.108, 1.202)	0.10
Drinker subgroup	3.44	1.235	0.454	(0.345, 2.125)	0.0065
Non-drinker subgroup	0.66	-0.412	0.526	(-1.444, 0.619)	0.43
Interaction (ratio of OR)	5.19	1.647	0.695	(0.285, 3.001)	0.018

Odds ratio (ranitidine/lansoprazole)

Interaction is ratio (drinker/non-drinker) of the odds ratios (ranitidine/lansoprazole)

Note: adjusted treatment effect is from logistic model and is not the simple weighted estimate of Λ

Analyses produced using SAS procedure GENMOD (SAS, 1996)

In Table 4.IV, estimates and confidence intervals are provided for all parameters including the within subgroup treatment differences and the interaction between treatment and drinking status. In addition to being more comprehensive, these (logistic) analyses are

now coherent with all parameters estimated using the same formulation (the log odds scale).

The unadjusted (1.75) and adjusted estimates (1.72) of the overall treatment difference are similar but notably less than the odds ratio of 3 specified for the alternative hypothesis. Both treatment groups contained a similar percentage of drinkers – that is, 60% (35/80) in the ranitidine group and 55% (43/78) in the lansoprazole group. (As shown in Chapter Three, Section 3.6, underestimation of the odds ratio would be expected under the condition of perfect balance, if an influential factor were excluded from the model. However in this case the adjusted OR is actually numerically smaller. Also note that the standard error (SE) for the adjusted analysis is numerically larger - a topic for further investigation in Chapter Five.) In terms of separate subgroup analyses, the treatment difference for drinkers is actually statistically significant (p=0.0065) while the corresponding difference for non-drinkers is not significant (p=0.43). As was highlighted in Chapter Two (Section 2.4.3), the chances of observing inconsistent subgroup results is reasonably high when the overall p-value is small, so this apparent inconsistency in terms of significance should not in isolation be viewed as a particularly worrisome finding. The SE in each subgroup is, as would be expected from Section 4.2.2, larger than the unadjusted treatment difference SE but smaller than the SE for the interaction.

The striking feature of the data is the magnitude of the estimate of the interaction parameter. Relative to the unadjusted estimate of the treatment difference, the interaction estimate is almost three times as large, and is also 1.7 times larger than the treatment difference that the study was designed to detect. (When using the odds formulation, the interaction parameter is not an odds ratio *per se*, but is in fact the ratio of the odds ratios.). As described in Chapter Two (Section 2.4.4), perhaps interactions analyses that are either exploratory or unplanned (which these most likely were) should be subject to an

adjustment for multiplicity. In this case, since eight factors were considered then both the Bonferroni and Holm (1979) corrections would compare the smallest p-value with α/k - that is, compare 0.018 with (0.05/8) =0.00625 – in which case the interaction would not now be statistically significant at the 5% level.

At this stage it is interesting to investigate the interaction further by applying the standard LR and SR tests as described in Section 4.2.3 to the data in Table 4.III. The results are reported in Table 4.V while the corresponding critical values for the tests are given in Table 4.VI.

Table 4.V: LR and SR test statistics for the investigation of treatment by baseline drinking status interaction

Statistic	Log OR (ranitidine/lansoprazole)	
	Drinker	Non-drinker
D_j	+1.235	-0.412
s_j	0.454	0.526
D_j^2 / s_j^2	7.400	0.613
D_j / s_j	2.720	-0.783
Q^-	7.400	
Q^+	0.613	
$\max\{D_j / s_j\}$	2.720	
$\min\{D_j / s_j\}$	-0.783	

Recall that with the LR test, the null hypothesis of no qualitative interaction is rejected if $\min(Q^+, Q^-) > C_{2\alpha}$ so given that 0.613<2.71, the null hypothesis is not rejected at the 5% level (p>0.20). Similarly the null hypothesis is not rejected at the 5% level when applying the SR test, since although 2.720 is greater than 1.64, -0.783 is also greater than -1.64. (Recall that rejection of the null hypothesis requires both $\max\{D_j / \sigma_j\} > C'_{2\alpha}$ and $\min\{D_j / \sigma_j\} < -C'_{2\alpha}$.)

In terms of the one sided null hypothesis ($H_{01} : \Delta \in 0^+$) that lansoprazole is at least as good as ranitidine in both subgroups, then for the LR test, this is rejected if $Q^+ > C_{1\alpha}$. In this case 0.613<4.23 so the null hypothesis is not rejected at the 5% level (p>0.20). Similarly for the SR test, H_{01} is rejected if $\min\{D_j / \sigma_j\} < -C'_{1\alpha}$ and since -0.783 is greater than -1.95, the null is not rejected. However for both the SR and LR tests, the one-sided hypothesis ($H_{01} : \Delta \in 0^-$) that ranitidine is at least as good as lansoprazole, in both subgroups, is rejected at the 5% level of significance (p<0.025).

Table 4.VI: Selected critical values for two-sided and one-sided LR and SR tests

Test	Critical value	Significance level				
		0.20	0.10	0.05	0.025	0.001
LR test	$C_{2\alpha}$	0.71	1.64	2.71	3.84	9.55
	$C_{1\alpha}$	1.73	2.95	4.23	5.54	11.76
SR test	$C'_{2\alpha}$	0.84	1.28	1.64	1.96	3.09
	$C'_{1\alpha}$	1.25	1.63	1.95	2.24	3.29

Critical values for LR and SR tests reproduced from Table 1, Gail and Simon (1985) and Table 1, Piantadosi and Gail (1993) respectively

These additional analyses provide further insight to the data and suggest that there is not strong evidence to support the presence of a qualitative interaction. Importantly, it has also been shown that ranitidine is not at least as good as lansprazole in both subgroups and therefore that lansoprazole is better than ranitidine in at least one subgroup. Furthermore, the evidence does not support the view that ranitidine is better than lansoprazole in any subgroup. In this respect a rationale dosing strategy for future patients might be to use lansoprazole 30mg rather than ranitidine 300mg in the treatment of oesophageal stricture, regardless of drinking status (since lansoprazole will not be worse). However these analyses make no attempt to estimate the magnitude of the interaction effect directly.

Now, consider the specification of interaction margins to aid interpretation. Since this study was designed to detect a difference (δ) of 25% between the treatments, then this is

the starting point for margin specification, and using the odds formulation, this reduction in the re-dilation percentage from 50% to 25% corresponds to an odds ratio of 3 (ranitidine:lansoprazole) - which when transformed to the log scale gives 1.099. Following on from Section 4.4, four margins will be considered. First of all, three based on δ - which in increasing order of width are: $(-\delta/2,+\delta/2)$; $(-\delta,+\delta)$; and $(-2\delta,+2\delta)$. The fourth margin is based on the observed $\hat{\Lambda}$ - that is, $(-2\hat{\Lambda},+2\hat{\Lambda})$. Now, in terms of constructing margins on the odds scale, it is the log OR that is used. For instance, to construct the margins $(-\delta/2,+\delta/2)$, the log OR of 1.099 is divided by 2 – that is, ($\log_e$ OR)/2 – to give ± 0.549. Since $\hat{\Lambda} = 0.411$ on the log scale – then the corresponding margin $(-2\hat{\Lambda},+2\hat{\Lambda})$ is simply ± 0.823. These margins are given in the left hand column of Table 4.VII below and it is worth noting that the values (± 2.197) for the margin $(-2\hat{\Lambda},+2\hat{\Lambda})$ equate to a ratio of the odds ratios of 9.

The next step in the process is to compare the estimated 95% confidence limits for the interaction on the log odds scale – that is, (0.285, 3.001) – with these margins. It is clear that although the lower 95% confidence limit is greater than zero, it is less than the upper margin in all four cases. In this respect the range of plausible values for the interaction includes values that are consistent with an interaction that is not clinically relevant using all four criteria specified (that is, all four sets of margins). However it is also worth noting that the estimate of the interaction itself is greater than the upper value for three out of the four margins.

Now, consider the application of the proposed Bayesian approach using the four margins defined above. First consider the case where a neutral prior is required and in this regard a Be(0,0) reference prior has been selected for each treatment by factor combination, which

is uniform for the log odds. The results are shown in Table 4.VII (reference priors column) where the probabilities quoted are posterior probabilities that the true interaction difference lies within the specified margins.

Table 4.VII: Posterior probabilities within a range of margins for the interaction parameter (treatment by drinking status) with reference and informative priors

Margins		Posterior probability contained within margins		
		Reference priors	Informative priors	
			5/combination	10/combination
$(-\delta/2,+\delta/2)$	± 0.549	0.056	0.081	0.108
$(-\delta,+\delta)$	± 1.099	0.215	0.293	0.370
$(-2\delta,+2\delta)$	± 2.197	0.786	0.874	0.928
$(-2\hat{\Lambda},+2\hat{\Lambda})$	± 0.823	0.118	0.166	0.217

The posterior probability for the margins $\leq |\delta|$ points to a high proportion of the interaction's posterior distribution lying outside the limits of the range. This would suggest that the interaction is clinically relevant in relation to the treatment difference that the study was initiated to detect. In contrast the margins $(-2\delta,+2\delta)$ include almost 80% of the posterior probability, although perhaps these are too wide from a practical standpoint since they correspond to 9 for the ratio of the odds ratios. Given that this study was powered to detect a difference between percentages of 25%, margins of double the magnitude are perhaps uninformative in this respect. The *a posteriori* margin $(-2\hat{\Lambda},+2\hat{\Lambda})$ contains only around 12% of the posterior distribution and again points to a clinically relevant interaction.

The question then arises as to what informative prior distributions might have plausibly been used for this study? Clearly the authors did not expect to observe an interaction between baseline drinking status and treatment although we know that the study was designed to detect a difference of 25%. In this respect one could argue that if asked before commencing the study the expected response percentages would have been 50% for both

combinations involving ranitidine and 25% for both combinations involving lansoprazole. However, the paper also references the results of a similar 12 month comparison of a different proton pump inhibitor (omeprazole) against a lower dose of ranitidine (that is, 150mg versus 300mg). These data were published in 1994 and showed re-dilatation percentages of 30% (43/143) for omeprazole (same class of drug as lansoprazole) and 46% (66/143) for ranitidine. One could argue therefore that it is more credible to think that the priors would have been based on these data. Clearly any attempt to retrospectively assign priors is a contradiction in terms, but for illustrative purposes assume that individual informative priors were specified prospectively for each treatment by factor combination. Assume further that the researcher expects the response percentage to be 40% for ranitidine (since a higher dose of ranitidine is expected to reduce the re-dilatation percentage) and 30% for lansoprazole (on the basis that the same class of compound might produce a similar effect). The next question is then what weight they would have given to these prior beliefs in terms of the number of observations? Assume that that researcher assigns priors $(\gamma_{ij}/(\gamma_{ij}+\varphi_{ij}))$ of 2/5 for each ranitidine combination and 1.5/5 (=30%) for each lansoprazole combination. (Note that integer values are not required.) In each case, no interaction with drinking status is expected and the weight given to each of the four priors – in terms of five subjects' worth of data – is the same. Essentially the researcher has added in 20 observations worth of data. Since this exercise is purely for illustrative purposes, consider a further case where the researcher is more certain with respect to their estimates and considers each to be worthy of ten subjects' data. In this case the respective priors for ranitidine and lansoprazole are now 4/10 and 3/10. The impact of applying these two sets of priors is shown in Table 4.VII above.

As would be expected, when the researchers applies priors with the expectation that there is no interaction (ratio of odds ratios equals one) then a greater proportion of the posterior

distribution will be squeezed between the margins (which are all symmetric around zero on the logarithmic scale). Furthermore, the greater certainty attached to the priors, the higher proportion of the posterior distribution is contained within the margins as can be seen when comparing the ten subjects per combination priors with the five subjects per combination priors. In the former case, although 40 subjects' worth of data has been included (weighed) with observed data from 158 subjects, the overall impression has not changes markedly for the margins, and in particular those $\leq |\delta|$. That is, a substantial proportion of the posterior distribution lies outside the margins.

As described in Section 4.5.4.3, an alternative approach to the analysis could be to use MCMC methods – specifically Gibbs sampling in BUGS. This approach is more flexible than the algebraic solution presented above and the researcher would not be restricted to formulating priors in terms of the beta distribution (the conjugate) for the binomial likelihood. With the algebraic solution proposed, the approximation to the Normal distribution has been used and as such there is no requirement to specify prior distributions for the variance parameters. In this respect the nuisance parameters are assumed known. BUGS software is flexible in this respect and there is no requirement to include prior information for the variance - although of course it is an option. (A full Bayesian solution requires prior distributions on both effects and variances of effects.) BUGS also allows for different model parameterizations such that it would be possible to formulate the model in terms on priors for each treatment by factor combination (including priors on each variance), or using a generalised linear model formulation similar to Simon and Freeman (1997) as described in Section 4.5.2. Other parameterizations would also be possible –for instance, a parameterization that specified priors for the treatment difference within each subgroup.

So in conclusion, one statistically significant interaction was observed when eight exploratory analyses were undertaken, but once an adjustment for multiplicity was performed the interaction was not significant at the 5% level. Although this interaction was substantial in magnitude – 1.7 times as large as the difference the trial was designed to detect – the hypothesis of no qualitative interaction was not rejected with the LR or SR tests. The one-sided LR and SR test pointed to lansoprazole being at least as good as ranitidine in both subgroups and better in at least one. As such with regard to the future treatment of patients, lansoprazole would be preferred to ranitidine – since regardless of drinking status it would be no worse – despite an overall significance level of p=0.093. (Note that other endpoints were also analysed and in particular statistically significant benefits were observed with lansoprazole in comparison with ranitidine in terms of reducing dysphagia (a symptom of the disease) and in controlling the concurrent condition of reflux oesophagitis.) Clearly the study was overly ambitious given the results from a previous study and was most likely under powered for a more modest - but more realistic - treatment difference. Applying a range of margins and the confidence interval for the interaction parameter confirmed that the data could be consistent with an interaction that was not clinically relevant. However the Bayesian approach pointed to a high proportion of the posterior distribution residing outside the margins $\leq |\delta|$. This was the case even when informative (but retrospectively chosen) priors were applied that were centred on no interaction being present. As such, a definitive conclusion is not possible and the recommendation would be that further data are required to shed light on the issue. For instance, one could investigate the pharmacokinetics (PK) of the two treatments with and without alcohol and also re-visit any studies that investigated the PK profile under various levels of hepatic function.

4.7 DISCUSSION

From a regulatory perspective, although the investigation of interactions is viewed primarily as an exploratory procedure, consistency of the treatment effect is clearly an important consideration when determining whether there is convincing evidence of a clinically useful effect. It is surprising therefore that given the limitations of the simple hypothesis testing approach identified in the guidelines, that estimation of the interaction parameter receives no attention. It is clear that quantitative interactions can only be ruled out if it can be demonstrated that a qualitative interaction is present, and in this respect estimation and the use of confidence intervals against pre-defined margins is an obvious route to clinical interpretation (with support, of course, from within subgroup estimates and associated confidence intervals). Indeed by stressing that all findings from interaction analyses should be treated with caution – both significant and non significant results – the regulatory authorities have implicitly introduced subjectivity and this perhaps marries poorly with their requirement for convincing evidence.

It should also be noted that in regulatory submissions, the requirement is not only to evaluate consistency of effect for confirmatory studies individually but also to evaluate consistency across all these studies combined to determine the level of support for the proposed dose schedule across specific subgroups. In regulatory terminology, the evaluations can be classified into one of three groups: drug-demographic, drug-drug or drug-disease interactions, and in the latter two cases the interaction of treatment with concomitant medications and concomitant diseases is explored. Presumably the rationale for this requirement - essentially a retrospective evaluation on the phase III database - is that the power to detect treatment by factor interactions will increase by combining studies. Again, however, power is irrelevant without consideration of what constitutes a clinically relevant difference.

As a first step the statistical guidelines could be modified to encourage the use of statistical procedures such as the LR and SR tests, and to promote greater thought in the area of *a priori* quantification of clinically meaningful interactions. In particular estimation and confidence intervals should be encouraged for both individual studies and when studies are combined. However given the current leaning towards subjectivity in assessment, the guidelines could do worse than advance formal Bayesian thinking in this area.

In Chapter Five, the thesis switches tack to focus upon a controversial area of drug development – that is, therapeutic equivalence and non-inferiority – where margin specification is fundamental to the methodological approach. The role of sub-populations will be investigated and the impact of adjustment for factors that influence outcome. It will be shown that not unlike treatment by subgroup interactions, the regulatory guidance in this area is also somewhat lacking and at times incoherent.

CHAPTER FIVE: THERAPEUTIC EQUIVALENCE: FALLACIES AND FALSIFICATION

Auntie Millie
Ran willy-nilly
When her legs, they did recede
And so they rubbed on medicinal compound
And now they call her Millipede.

5.1 INTRODUCTION

Although much has been published on the topics of equivalence and non-inferiority – including dedicated sections in recent regulatory guidelines, such as ICH E9 (1998), ICH E10 (2000) and CPMP/EWP/482/99 (2000) – some areas have been neglected and a number of common misconceptions prevail. The aim of this chapter is to highlight some of these areas and to challenge some of the practices employed - particularly in relation to constructing sub-populations and investigating the influence of prognostic factors.

Studies designed specifically to demonstrate therapeutic equivalence are a relatively recent development in pharmaceutical research. Early papers on the subject appeared in the late 1970s – Dunnett and Gent (1977), for instance – and numerous methodological papers have been published since, as the approaches employed for bioequivalence were adapted to the therapeutic setting. In particular the concept that the confidence interval for a treatment difference be wholly confined within a pre-specified equivalence range transferred readily from bioeqivalence to therapeutic equivalence.

The typical bioequivalence study compares a generic copy of a standard formulation of a drug with the standard formulation itself using a cross-over design in relatively few subjects. Since ostensibly the same drug is being compared, one would expect the same

therapeutic effect, but instead of evaluating clinical efficacy directly, these studies - which are typically conducted in healthy volunteers - use summary measures of the amount of drug present in blood through time (pharmacokinetic parameters) as indirect measures or surrogates of clinical efficacy.

Simply stated, if bioequivalence between the standard formulation and generic copy is demonstrated, then therapeutic equivalence is assumed by implication. The methods used for bioequivalence can also be applied to compare the same formulation of a drug in different settings such as with or without food (food interaction studies) or in between subject comparisons when evaluating the impact of varying degrees of renal or hepatic dysfunction. A detailed description of the issues surrounding bioequivalence is provided in European (CPMP/EWP/QWP/1401/98) and US regulatory guidelines (FDA, 2000 and FDA, 2001).

Therapeutic equivalence offers unique challenges, however. In contrast to bioequivalence, the effects of different drugs are usually compared and the primary endpoint is typically a direct measure of therapeutic benefit. Equivalence methodology has also been used to compare different formulations of the same drug – for instance the statutory requirement to replace CFCs in inhaled asthma treatments led to studies being conducted for a range of drugs. However opportunities for strict equivalence studies tend to be limited and pharmaceutical research objectives are more often directed towards the one-sided case where the conclusion "at least as good as the reference treatment" is sought. This is referred to as non-inferiority and will form the main focus of this chapter.

Both equivalence and non-inferiority studies are typically large confirmatory trials that employ a randomised parallel group design with an active control group. Some trials also

include a placebo control group - where ethically feasible - to both validate the study and to demonstrate superiority of the test treatment to placebo. These trials that include both active and placebo controls are most likely to be conducted in less severe chronic disease areas. However, non-inferiority methods are frequently used in more serious, acute and sometimes life threatening indications such as oncology and infectious diseases. In these cases the primary endpoint is most often a binary outcome measure and in some cases the trial objective will incorporate a time to event evaluation. With this in mind, binary outcomes will be the central feature of this chapter when issues relating to non-inferiority are discussed.

Section 5.2 of this chapter provides a general overview of therapeutic equivalence methodology. The design and analysis of these studies are discussed and compared with superiority trials. In Section 5.3, consideration is given to the different model formulations that may be employed for the comparison of binary outcomes. In particular, the standard difference in proportions is compared with the odds ratio as a measure of the relative treatment effect. In Section 5.4, the choice of the most appropriate subject population to employ is discussed and the conservative nature of the per protocol population is questioned. Strategies for validating equivalence trials will also be appraised here. In Section 5.5, covariate adjustment will be considered and in particular the behaviour of the odds (logistic) model will be examined. Finally, in Section 5.6, consideration will be given to sample size calculation and the perception that non-inferiority studies require a greater number of subjects in comparison with superiority studies. The inclusion of covariate information will also be discussed. In all sections, reference will be made, where applicable, to the current regulatory guidance that exists for drug development.

5.2 AN OVERVIEW OF EQUIVALENCE METHODOLOGY

Bristol (1999) provides a historical review of the development of equivalence methodology while Senn (1991) addresses the philosophical issues. In 2000, the Committee for Proprietary Medicinal Products (CPMP) produced a specific Points to Consider document (CPMP/EWP/482/99) which gives details of many of the practical issues associated with therapeutic equivalence and non-inferiority with an emphasis on estimation rather than hypothesis testing.

The methodologies for both equivalence and non-inferiority revolve round the pre-specification of equivalence margins ($-\delta_1$, $+\delta_2$) that are selected on the basis of defining the largest difference between test and reference treatments that would be clinically acceptable, and against these margins, plausible values for the true treatment difference are judged. Although these margins are commonly symmetric with respect to zero, it is noted that asymmetric margins may be appropriate in certain circumstances.

For equivalence, the formal hypothesis testing formulation involves the establishment of two sets of one-sided hypotheses. That is,

$$H_{01}: \mu_{test} - \mu_{reference} \leq -\delta_1 \text{ and } H_{02}: \mu_{test} - \mu_{reference} \geq +\delta_2,$$

against the corresponding alternatives,

$$H_{A1}: \mu_{test} - \mu_{reference} > -\delta_1 \text{ and } H_{A2}: \mu_{test} - \mu_{reference} < +\delta_2.$$

It follows that the null hypothesis (H_{01} or H_{02}) can be rejected at the α level in favour of the alternative (H_{A1} and H_{A2}) if simultaneously both H_{01} and H_{02} are rejected at the same α level. For non-inferiority, the approach reduces to a single one-sided test of the null hypothesis at the selected α level. That is,

$$H_0: \mu_{test} - \mu_{reference} \leq -\delta_1,$$

against the alternative,

$$H_A: \mu_{test} - \mu_{reference} > -\delta_1.$$

Recall that the conventional hypothesis testing formulation for superiority controls the probability of incorrectly rejecting the null hypothesis of equality (H_0: $\delta = 0$) whereas for equivalence or non-inferiority it is the probability of incorrectly accepting the (alternative) hypothesis of equivalence or non-inferiority (which includes equality) which is controlled.

The direct use of confidence intervals provides an alternative approach. Specifically, confidence limits can be used to provide boundaries for the plausible treatment differences since they are taken to represent quantitative estimates of the minimum and maximum estimated effects of a test treatment relative to a reference. Confidence intervals are operationally equivalent (ICH E9, 1998) to the use of one-sided tests and also provide a simple and effective way of communicating the results. A detailed comparison of the main approaches is given in Senn (2001).

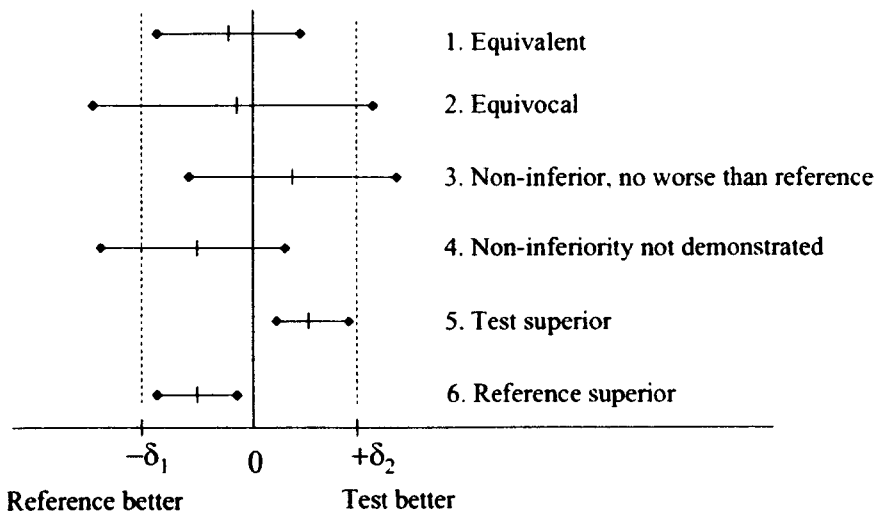
For equivalence and non-inferiority studies it is recommended (ICH E9, 1998; CPMP/EWP/482/99, 2000) that the α (or type I) error is set to 2.5%, leading in both cases to 95% confidence intervals. This approach promotes consistency with the general approach of estimating treatment effects in superiority studies, although note that in bioequivalence, 90% confidence intervals are the established standard. To maintain consistency with the referenced regulatory guidelines, the remainder of this chapter will now focus on the confidence interval interpretation of the problem.

Figure 5.1 illustrates how confidence limits for an estimated treatment difference – in association with equivalence margins - can be used to provide an appropriate clinical

interpretation of the trial results. Note that in the figure a treatment difference of zero could refer to either a difference in percentages or a difference in the log odds (equating to an odds ratio of 1) for binary outcomes.

For equivalence, interest is focused on both the upper and lower confidence limits and equivalence can only be declared if the entire confidence interval is contained within the margins as shown in Case 1. Even if the estimated treatment difference is zero, a conclusion of equivalence can not be drawn unless this condition holds. Case 2 provides an example where the conclusion is ambiguous; i.e. the treatment difference is equivocal.

Figure 5.1. Basic principles underlying confidence interval approach to equivalence/non-inferiority



In contrast, for non-inferiority, the primary interest is in the lower confidence limit only. To claim non inferiority the only requirement is that the lower confidence limit is greater than the lower equivalence margin as shown in Case 3. If the lower limit is below the margin then larger differences in favour of the reference treatment can not be ruled out and

non-inferiority has not been demonstrated (Case 4). Cases 5 and 6 represent apparent contradictions. In both examples the whole CI is contained within the margins yet both would yield a statistically significant p-value against a traditional null hypothesis of no difference. Now, for bioequivalence, there are well-established equivalence margins (CPMP/EWP/QWP/1401/98, 2001; FDA, 2000; FDA, 2001) and these are rigidly applied by the regulatory authorities. If a confidence interval were to exclude the point of no difference then this would be of little concern and bioequivalence would still be accepted. However for therapeutic equivalence and non-inferiority, margin specification is less well developed and in particular if the margins applied were wide, then acceptance of the claim may prove more difficult to achieve for Case 6. Taking the argument a step further, if the endpoint were survival or response in a major life threatening disease it might be unreasonable to conclude anything but inferiority in Case 6 while in Case 5 it would seem reasonable to claim superiority of the test treatment (Röhmel, 2001). The argument here is that with survival the risk/benefit is usually clear and even a small but precise difference is clinically relevant. Situations like cases 5 and 6 rarely occur in practice.

In relation to Case 5, it is considered acceptable to first test for non-inferiority and if this is demonstrated to test for superiority using a null hypothesis of no treatment difference (CPMP/EWP/482/99, 2000). Due to the closed test nature of the testing procedure, there is no multiplicity issue (Morikawa and Yoshida, 1995) although one should be cautious of unexpected positive findings. The reverse situation of interpreting a superiority trial as a non-inferiority one is not quite so straightforward since it is unlikely that a non-inferiority margin would have been pre-specified to enable an objective determination to be made.

For these switching strategies to be valid there must also be consistency with regard to both subject analysis populations and confidence interval coverage. For instance, if a

reduced per protocol type population were used for non-inferiority then the closed testing procedure would not hold if an intent-to-treat type population were subsequently used for superiority. Similarly, if would be inappropriate to use a 95% confidence interval for superiority but switch to a lower 90% confident limit for non-inferiority (since historically 90% was used for many non-inferiority protocols). As such, full details of switching strategies should always be pre-specified in the study protocol.

Finally in this overview, it is important to highlight the issue of trial validity. Senn (1991) coined the phrase *competence* to describe *the ability of the trial to detect a difference in treatments if it exists*. (The term 'assay sensitivity' is also sometimes used to describe this attribute [ICH E10, 2000].) The dilemma faced with equivalence is that if a study achieves its objective - that the two treatments are equivalent - it is not possible to show that the study was competent. Note that randomisation and blinding only strengthen a study conclusion if that conclusion is that the treatments differ since no knowledge of the treatment allocation is required to create a conclusion of equivalence (Senn, 1991); this can be achieved by simply assigning a random response to all subjects. To address the issue of competence, some designs have incorporated a placebo arm where this is considered ethically acceptable; by showing superiority of the test treatment against placebo a degree of competency is shown and as such the comparison of test with reference is validated.

5.3 SPECIFICATION OF EQUIVALENCE MARGINS

The primary aim of this section is to demonstrate that the odds ratio is the most rational measure for assessing therapeutic equivalence and non-inferiority for binary outcomes and that there are clear advantages to expressing margins in terms of the odds ratio. A useful starting point is to review the current regulatory stance with regard to these margins.

Röhmel (1998) has previously reviewed non-inferiority criteria and in particular two anti-infective guidelines (FDA, 1997; CPMP/EWP/558/95, 1997) that attempt to provide tangible guidance for binary outcome measures. (Anti-infectives is an area where non-inferiority methodology has been widely employed in recent years following the introduction of a number of specific regulatory documents. It is particularly appropriate in this indication since placebo controlled studies are considered unethical and response rates with current treatments are relatively high leaving modest scope for improvement.) Non-inferiority criteria dependent upon the highest of the two observed response percentages being compared were initially given in a draft guideline issued by the FDA (1997) although these criteria were withdrawn in 2002. Specifically, a margin of 10% (that is, ten percentage points) was reserved for responses of 90% or higher, while 15% was specified for responses of at least 80% but less than 90%. Finally, a 20% margin was to be applied to responses <80%. A striking feature of this convention was that a more stringent criterion was created for the test treatment that performed well in an individual clinical trial. For instance, if a response of 89% were observed for the reference treatment then a response of 91% for the test would require a margin of 10% for the lower confidence limit of the difference whereas a response of 88% would require a margin of 15%.

In contrast, in a guideline issued by the CPMP (CPMP/EWP/558/95, 1997), a fixed non-inferiority margin of about10% for many non serious infections - regardless of the percentage response for the reference treatment - is specified. The statement that for very high responses a smaller margin will be needed, qualifies this rule.

Röhmel (1998) criticises both these anti-infective rules and argues that, from both a statistical and clinical perspective, the margin should vary with the response of the

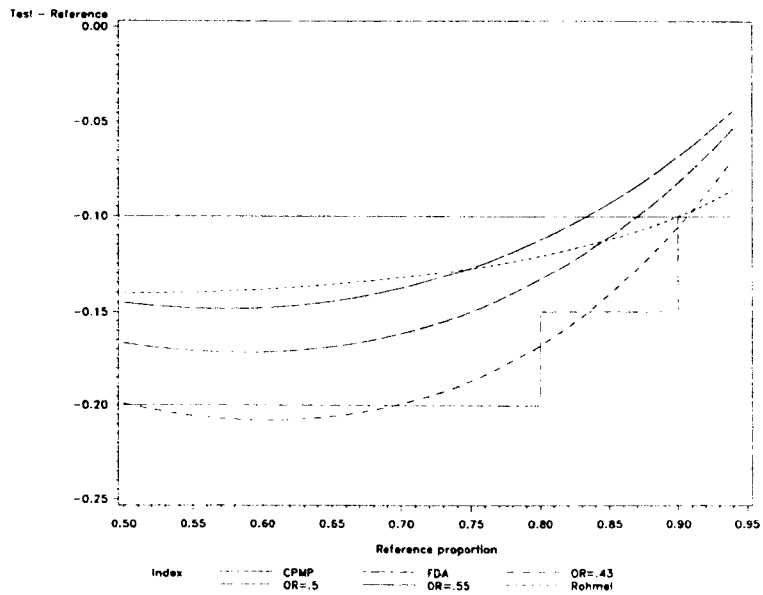
reference treatment, but that unlike the FDA criteria, the transition should be smooth rather than a step function. Röhmel (2001) subsequently proposed a margin of -0.223 $\sqrt[3]{p(1-p)}$, where p is the observed proportion for the reference treatment. However a fixed margin for an odds ratio provides the most obvious solution to both problems.

Recall that the odds ratio provides a measure of the difference between treatments that is stable over a wide range of conditions such that the overall proportion is arbitrary (Cox, 1970). (This is not the case for the difference between proportions that is restricted to a range of values for the individual proportions generated from each treatment.). As such, it is possible to specify one value for the lower margin, 0.5 say, that is applicable to broad range of reference responses. In this respect it is similar in philosophy to the bioequivalence rule specified by the regulatory authorities (CPMP/EWP/QWP/1401/98, 2001; FDA, 2000; FDA, 2001) that requires the 90% confidence interval for the ratio of the area under the curve for the reference drug to the test drug to be wholly contained within the boundary 0.8 to 1.25. In the case of therapeutic equivalence or non-inferiority, a constant odds ratio margin would correspond to smaller differences in proportions as the reference treatment proportion approached either 0 or 1. This is consistent with the philosophy of both the FDA and CPMP rules and with usual clinical trials design practice. Furthermore, an odds ratio margin would provide a smooth transition when mapped to the difference in proportions as the reference proportion changed and would avoid the step like function of the FDA rule. As a result, the reductions in power that occur at the FDA defined thresholds would be avoided in cases where the highest observed response was higher than anticipated.

Figure 5.2 shows the CPMP and FDA rules for reference proportions ranging from 0.5 to 0.9. For simplicity it is assumed that for the FDA rule, the reference proportion is at least

as high as that for the test treatment. Superimposed onto the figure is an odds ratio lower margin of 0.5 - translated into a difference in proportions for each reference proportion - to illustrate the features described earlier. A simple value of 0.5 appears to embody the philosophies of both the FDA and CPMP rules and for most reference proportions provides a compromise between the two in terms of the difference in proportions. Against the backdrop of the FDA rule, Tu (1998) has proposed a value of 0.43 and although equally valid, it can be seen from Figure 2 that Tu's margin would be less strict than both regulatory margins for some reference proportions and would always be less strict than 0.5. It does however follow the route of the FDA rule and would provide a simple alternative. In the more general context, Senn (2000) has suggested 0.55 (as illustrated in Figure 5.2). Senn's suggestion was made on the basis of a regulatory agency that may wish to ensure a maximum possible inferiority of 0.15 when considering the difference in proportions. However in the context of anti-infectives, it would be more stringent than both the FDA and CPMP rules at fairly modest reference response rates and beyond (>0.833). Röhmel's formula (Röhmel, 2001) is based on the observed reference proportion but appears to work well. It intersects the FDA and CPMP rules when the observed reference proportion is 0.90 and is contained within a narrower band when compared with both the odds ratio rules and the FDA rule. However since it uses the difference scale it would be likely to lead in many cases to margins which appeared somewhat contrived. For instance an observed reference proportion of 0.80 leads to a margin of 0.121 (12.1%). The advantage of the odds ratio is that, on its chosen scale, the margin is constant.

Figure 5.2. Non-inferiority margin for difference in proportions for CPMP, FDA, odds ratio and Röhmel rules



As far as the author is aware, no regulatory documents exist which propose margins for the odds ratio. Indeed, it is interesting to note that the CPMP points to consider document (CPMP/EWP/482/99, 2000), issued as recently as July 2000, provides no guidance whatsoever as to which margins should be adopted. It does however reference a concept paper (CPMP/EWP/2158/99, 1999) that expresses the intention of the CPMP to address this issue, although the concept paper does not itself constitute guidance. Indeed, the paper notes that the difficulties surrounding the issue mean that definitive guidance about how to choose non-inferiority margins may not be possible. It briefly describes some of the more common practices such as the use of *one half to one third of the established superiority of the comparator to placebo*, and it acknowledges that mortality studies represent a unique challenge. It also refers to anti-infective studies where the margin may *vary according to the expected response rate to the standard treatment.* In 2004, the CPMP finally replaced the concept paper with a draft *Points to consider on the choice of non-inferiority margin* (CPMP/EWP/2158/99 draft, 2004). However this document proved to be less prescriptive than had been earlier indicated and in many places ventured into new and, most likely, untested territory, and as a result has proved highly

controversial. Indeed in an apparent break with tradition, it even suggested that model formulation should depend upon the observed responses. For instance, for binary outcomes it proposed that the margin should be specified on both the odds ratio and difference in proportions scales with the most conservative chosen a once the data have been collected. Since model formulation is the first step in quantifying the primary hypothesis to be tested, this proposal appeared at best bizarre and it did not seem sensible for the null and alternate hypotheses to depend upon the observed data. (In fact as will be shown later in this chapter in relation to factor adjustment (and as illustrated in Chapter Four in terms of interactions), if a stratified design has been employed such that the primary analysis is also stratified (ICH E9, 1998), then there are important differences between the odds ratio and difference in proportions formulations.) In July 2005 the final document was issued as a guideline entitled *Guideline on the choice of the non-inferiority margin* (CPMP/EWP/2158/99, 2005) with this section on formulation switching deleted following comments received during the consultation period. (Note that the author's comments pertaining to this specific point were included as part of the PSI [Statisticians in the Pharmaceutical Industry] response to the CPMP.)

Others have suggested a margin *less than half the difference between active and placebo* (Phillips *et al*, 2000). However as highlighted earlier, the most recent guidance has brought a less prescriptive approach and methodological research (Rothmann *et al*, 2003) is now being directed towards techniques that aim to compare the estimated confidence interval for test minus reference with the confidence interval for the reference effect (based on a meta analysis of historical data comparing reference treatment to placebo). In effect a new two-stage hierarchical procedure has been introduced through the *Guideline on the choice of the non-inferiority margin* (CPMP/EWP/2158/99, 2005). The first step (a minimum requirement) is to establish that test treatment is more effective than placebo

while the second step (conditional on the fist step being passed) is to establish that a deficit, in terms of a pre-specified proportion of the reference treatment effect, is implausible. (Ideally the first step would be achieved by including a concurrent placebo control in the study design although as noted previously this is not always possible.) In this respect it has been highlighted that the fixed margin approach of Phillips *et al* (2000) does not guarantee that the first step is satisfied and as a result the guideline does not endorse this method.

Rothmann *et al* (2003) re-iterate the non prescriptive approach to margin specification and note that margins need to be quantified on a case by case basis. Specifically they describe the dual requirement for clinical judgment (in terms of determining the proportion of the reference effect to be retained to support a claim of non-inferiority) and statistical method (in terms of estimating the reference effect from historical data). As an illustration they note that if the reference effect is large (for instance the treatment of bacterial infection, lymphoma and leukaemia) then there would be a requirement to retain a high proportion of the reference effect and as such the margin would be chosen primarily on the basis of clinical judgment. In contrast if the effect of reference were small then the clinically relevant difference could conceivably be greater than the reference effect. In this case showing that the difference between test and reference was less than the clinically relevant difference would not actually demonstrate that the test treatment was effective.

Finally, it is important to note that a further advantage to using the odds ratio is that odds is easily incorporated into the generalised linear model framework of Nelder and Wedderburn (1972) through the logit transformation of the proportion. This leads to parameter values in the real plane rather than in the unit square (Cox, 1970). With the logistic regression model it is simple to adjust the estimate of the treatment effect for the

impact of one or more important covariates (although as will be shown later in Section 5.5, this process of conditioning is not quite so straightforward as one might think).

For completeness, the relative risk (ratio of proportions) and the number needed to treat (NNT, defined as the inverse of the difference in proportions) are discussed briefly. Lack of symmetry is a major disadvantage for the relative risk formulation while on the difference scale the margins become wider as the reference proportion approaches 1, which seems counterintuitive. For instance, if the lower margin for the relative risk were 0.5, then on a difference scale this would equate to a lower margin of 25% for a reference proportion of 0.5 but 45% for a reference proportion of 0.9. A further important consideration (Cox, 1970) is that if success and failure are interchanged, the difference between groups is altered for the relative risk whereas the odds ratio is invariant in this respect. Tu (1998) discusses the relative risk formulation in more detail but ultimately recommends against it. NNT is the average number of subjects needed to be treated with one treatment to achieve one additional positive response compared with another treatment and has proved to be a popular summary measure for some in evidence-based medicine. However it is severely limited in terms of statistical modelling as it is a non-monotonic function with singularity for zero treatment differences. This can lead to a disjoint confidence interval, for instance. Hutton (2000) provides a highly critical review of NNT stating that it *at best conveys only the same information as the difference in proportions.*

5.4 SUBJECT ANALYSIS POPULATIONS

5.4.1 The per protocol population

It has become increasingly common to exclude protocol violators (PV) from the primary subject population when analysing equivalence and non-inferiority trials. The dominance of the per protocol type population (PP) has to a large extent been driven by publications (Lewis and Machin, 1993) and guidelines (CPMP III/3630/92-EN, 1995; ICH E9, 1998;

CPMP/EWP/482/99, 2000) that state that intent to treat (ITT) type populations tend to dilute the estimated magnitude of any treatment difference. For binary outcomes, the argument for dilution presumably originates from the misclassification type models presented earlier in Chapter Two (Sub-section 2.3.2.) It is argued therefore that the exclusion of PV's produces a refined subject population that is more capable of distinguishing treatments – that is, is more sensitive to treatment differences. At the same time, these exclusions lead to an increase in the standard error of the estimated treatment difference leading to a conservative approach from a regulatory perspective. However, as PVs are typically determined not only on the basis of pre-randomisation information but also post-randomisation recordings, such as compliance, there is potential for the introduction of bias when these subjects are excluded since independence of a post-randomisation measurement in relation to the randomised treatment is uncertain. Indeed this is the very point that led to the pre-eminence of the ITT type populations in the reporting of superiority trials, since - over all randomisations - the treatments groups will be balanced, whereas for the PP type populations this will not necessarily be the case.

To examine the potential impact of employing PP-type populations for non-inferiority determination, an approach adapted from Choi and Lu (1995) has been adopted. Choi and Lu studied the impact of missing data under conditions where the mechanism for observing the missing data was not completely at random (Little and Rubin, 1987). For binary outcome data, the probability of a response being missing for a specific treatment group was assumed to be related to the population probability of response in that treatment group. A similar approach is applied here with the probability of a violation replacing the probability of a response being missing.

Let π_r be the probability of response (x = 1) in a population receiving a reference treatment. Similarly, let π_t refer to the corresponding probability of response for the test treatment. If θ_1 is the probability of a subject being a PV when the response would have been 1 and θ_0 is the probability of a subject being a PV when the response would have been 0, it follows that the probability of not being a PV in the reference group is given by:

$$\lambda_r = \pi_r (1 - \theta_1) + (1 - \pi_r)(1 - \theta_0) \qquad (1)$$

As such, the expectation of the sample proportion (p'_r) for the reference group excluding the PV's is given by:

$$E(p'_r) = \pi_r (1 - \theta_1)/\lambda_r$$

Similarly for the test group, $E(p'_t) = \pi_t (1 - \theta_1)/\lambda_t$

The expected difference between the test and reference groups in the PP population is then:

$$E(p'_t) - E(p'_r) = \pi_t (1 - \theta_1)/\lambda_t - \pi_r (1 - \theta_1)/\lambda_r$$

with variance

$$(1 - \theta_1)(1 - \theta_0)\, n^{-1} \{\pi_r (1 - \pi_r)/\lambda_r^3 + \pi_t (1 - \pi_t)/\lambda_t^3\}$$

where n is the number of patients in each treatment group in the full population without violators.

To illustrate the impact of these results on the evaluation of therapeutic non-inferiority Figure 5.3a compares the expected treatment differences and 95% confidence limits from a PP population with those from a full population (with no violators) for reference proportions ranging from 0.2 to 0.9 and for $(\pi_t - \pi_r) = -0.1$. (This effectively limits the proportions to between 0.1 and 0.9.) Figure 5.3b shows the corresponding cases when the difference in the population proportions is zero, that is, $(\pi_t - \pi_r) = 0$. In both figures the confidence limits have been constructed using the Normal approximation to the binomial

distribution, and the range of proportions has been selected such that this approximation remains valid.

Figure 5.3a. Difference in proportions for full data set and per protocol population: 200 subjects per treatment, ($\theta_1 = 0.1$; $\theta_0 = 0.4$; $\pi_t - \pi_r = -0.1$)

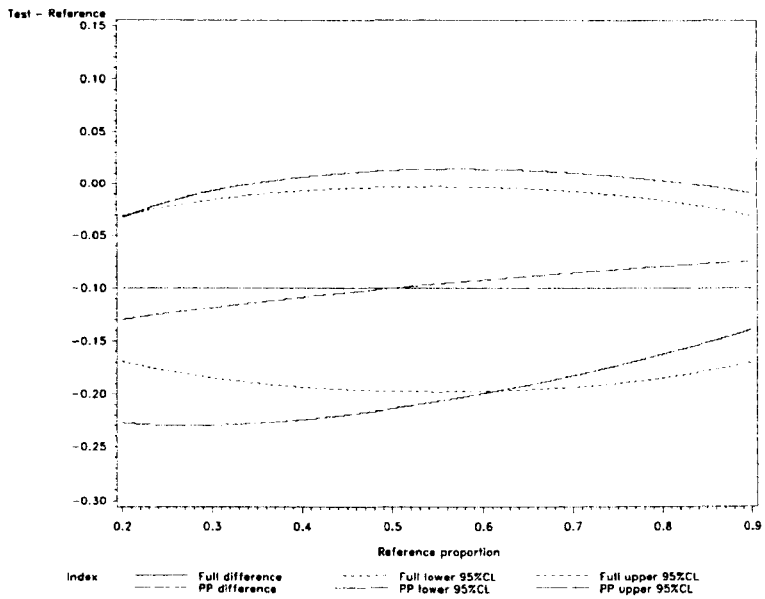


Figure 5.3b. Difference in proportions for full data set and per protocol population: 200 subjects per treatment, ($\theta_1 = 0.1$; $\theta_0 = 0.4$; $\pi_t - \pi_r = 0$)

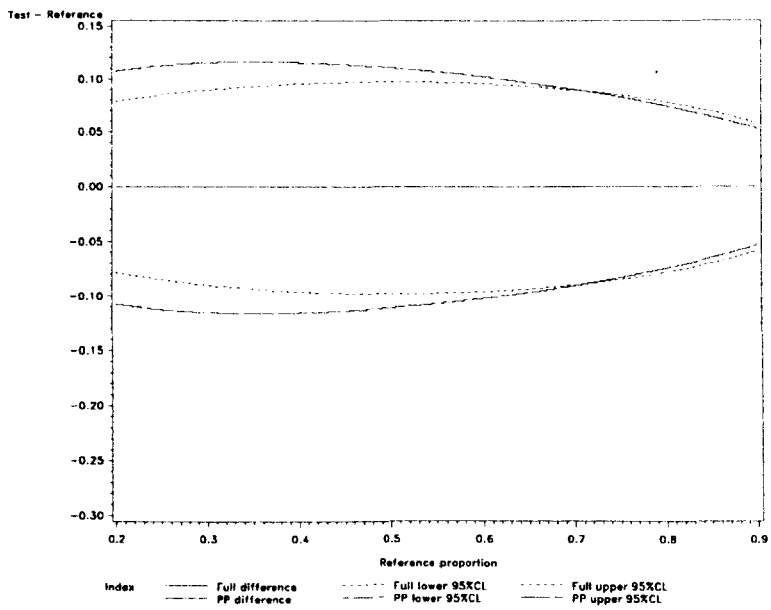


Figure 5.3a illustrates how bias can influence the estimate of the population treatment difference in both directions depending upon the reference proportion. In the case shown where $\theta_1 < \theta_0$, the difference between treatments is underestimated for high reference proportions but overestimated for low ones. When $\theta_1 > \theta_0$ the direction of the bias is reversed. However, as described earlier, the focal point for non-inferiority is the lower confidence limit. For the PP population, the value for the lower confidence limit is determined by three factors: potential bias in the estimate of the difference in proportions; the resulting number of subjects for each treatment; and the individual $p'(1 - p')$ terms for each treatment. The latter two factors influence the standard error of the estimate. As shown in Figure 5.3a, these factors can combine to produce an effect that is actually anti-conservative for the PP population. For reference proportions from around 0.6, the bias dominates the lower limit. The example chosen has 400 subjects in the full data set, but since bias is independent of the number of subjects whilst the standard error is inversely proportional to it, the relative importance of the bias will increase as the number of subjects increases. As such, if the number of subjects in the trial were to be increased then the cross-over point in Figure 5.3a would move towards the left and the region of anti-conservativeness would increase.

Figure 5.3b illustrates the expected finding that when there is no difference between the treatments the confidence interval for PP is generally wider, reflecting the reduced sample size. However, for high responses the influence of the terms $p'_t(1 - p'_t)$ and $p'_r(1 - p'_r)$ counteracts the reduction in subjects in both groups since $E(p'_t) > \pi_t$ and $E(p'_r) > \pi_r$. As $\theta_1 < \theta_0$, more potential failures than potential successes are excluded to form the PP population and as a result the expected observed proportions are higher in both treatments. When $\theta_1 > \theta_0$, a similar effect is observed but this time for the small population proportions.

Interestingly, with this particular formulation of the PV mechanism, the odds ratio for the PP population is unbiased when the treatment difference is not unity and its conservative nature is entirely due the reduction in subject numbers. Consider the odds ratio for the full population with no PV's:

$$\psi_T = \pi_t (1 - \pi_r) / \pi_r (1 - \pi_t)$$

The expectation of the odds ratio (ψ'_T) from the PP population is:

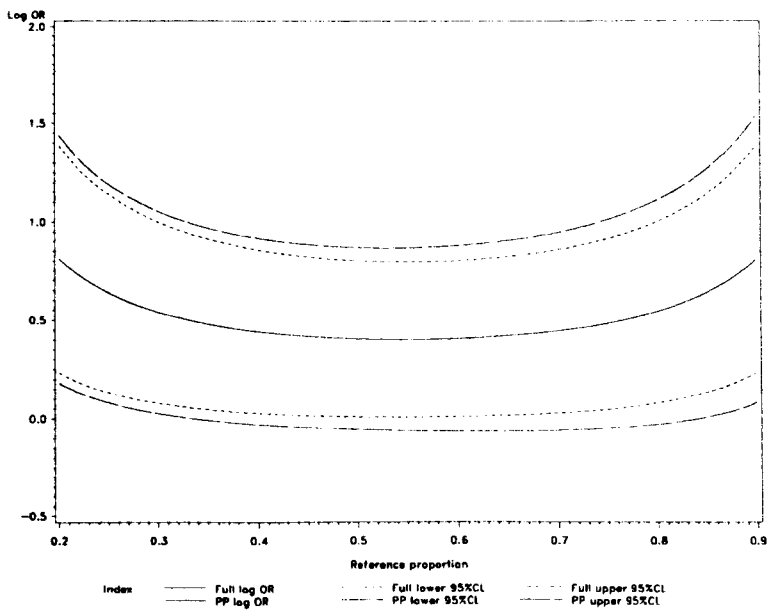
$$\pi_t [\lambda_r - (1 - \theta_1) \pi_r] / \pi_r [\lambda_t - (1 - \theta_1) \pi_t]$$

which gives, when substituting in λ_r and λ_t using expression (1),

$$\pi_t (1 - \pi_r)(1 - \theta_0) / \pi_r (1 - \pi_t)(1 - \theta_0) = \psi_T$$

This feature is illustrated in Figure 5.4 where the log odds ratio and associated 95% confidence limits for the PP and full populations are plotted against a range of reference proportions.

Figure 5.4. Log odds ratio for full data set and per protocol population: 200 subjects per treatment, ($\theta_1 = 0.1$; $\theta_0 = 0.4$; $\pi_t - \pi_r = -0.1$)



5.4.2 The eligible population

Now, consider the subgroup of eligible subjects. Subjects who do not have the disease (or in some cases, do not have the required severity of disease) cannot by definition demonstrate a response to treatment. For instance, an ineligible subject may be one with clinical signs and symptoms of a urinary tract infection in a trial of an anti-infective treatment but in whom no pathogens or bacteria are cultured from a urine sample. The determination of a bacteriological response is therefore impossible. A second example is a subject who treats a mild migraine attack where a positive outcome is defined as a reduction in headache severity from moderate or severe to none or mild. Note that in both examples described the proportion of ineligible cases may be non-trivial. It is somewhat inevitable therefore that the inclusion of these ineligible subjects will dilute the treatment difference.

However, as disease severity in both these cases is recorded prior to receiving blinded treatment, no bias in terms of the randomisation is introduced by their exclusion. This is the case even if the values are not known until some time after trial treatment has commenced – samples for microbiological testing, for instance.

As eligibility has been shown to be independent of response, the probability of a randomised subject being eligible in the reference group is now simply

$$\lambda_r = (1 - \theta_2),$$

where θ_2 is the probability of a subject being ineligible.

With this formulation, misclassification of response relates to the probability of a positive response only (that is, a response of 1) since one only needs to consider the case of what the response would have been had the subject had the disease.

As such, the estimate of π_r from the eligible population is unbiased and is given by

$$E(p''_r) = \pi_r (1 - \theta_2)/\lambda_r = \pi_r$$

Similarly for the test group, π_t

In contrast, the estimate of π_r from an ITT population - which now includes both $n\lambda_r$ eligible and $n(1 - \lambda_r)$ ineligible subjects – is:

$$E(p_r) = \pi_r (1 - \theta_2)$$

Similarly for π_t, $E(p_t) = \pi_t (1 - \theta_2)$

The estimate of the treatment difference $\pi_t - \pi_r$ is given by:

$$E(p_t) - E(p_r) = (\pi_t - \pi_r)(1 - \theta_2)$$

This is essentially a standard misclassification model that shows dilution of the treatment difference estimate in the presence of non-differential misclassification (Goldberg, 1975).

5.4.3 Supporting simulations

The next stage is to consider the practical implications of these observations by examining the probabilities of making incorrect decisions [p(non-inferior | inferior) and p(inferior | equivalent)] when a PP type population is used to compare two treatments with regard to response proportions. To do this, one needs to consider the probability distribution for the lower 95% confidence limit for the difference between the test and reference treatments, which itself is a random variable.

Table 5.I shows the results of some simulation exercises where the sample sizes for the treatments have been determined using the method of Makuch and Simon (1978) with a one-sided type I error of 2.5% and type II error of 20%. For illustration, consider the case where the reference proportion is 0.9 and the margin of non-inferiority is 0.1. When the test and reference proportions are identical, the power of the test is 86.7% for [$\theta_1 = 0.1, \theta_0 = 0.4$] but only 48.6% for [$\theta_1 = 0.4, \theta_0 = 0.1$]. The reason for this can be seen in Figure

5.3b. When $\theta_1<\theta_0$, the 95% CI is actually narrower for the PP population than for the full subject population when $\pi_r = 0.9$ since more potential failures are excluded than successes, increasing the observed proportions with the subsequent impact on the variance. This impact outweighs the reduction in the number of subjects. However, when $\theta_1>\theta_0$, more successes are excluded and both features work to increase the standard error of the estimate.

In contrast, when the test proportion is 0.8 (that is, inferior to reference) the type I error is 11.6% for [$\theta_1 = 0.1, \theta_0 = 0.4$] but 0.9% for [$\theta_1 = 0.4, \theta_0 = 0.1$]. In both cases bias is now a feature. When $\theta_1<\theta_0$ (Figure 5.3a), the bias is towards 0 which together with the smaller standard error leads to inflation of the type I error, whilst when $\theta_1>\theta_0$, the bias is in the opposite direction and when combined with the larger standard error reduces the chances of showing non-inferiority.

Table 5.I. Simulation 1: Impact on acceptance/rejection of non-inferiority when comparing full subject set to per protocol type populations for the difference in proportions

π_r	π_t	$\pi_t - \pi_r$	% non-inferior					N per treatment in Full Set
			Full	$\theta_1 = .1$, $\theta_0 = .2$	$\theta_1 = .1$, $\theta_0 = .4$	$\theta_1 = .2$, $\theta_0 = .1$	$\theta_1 = .4$, $\theta_0 = .1$	
.50	.40	-.10	2.56	2.46	2.78	2.72	3.42	393
	.50	0	79.00	73.64	69.58	73.96	69.66	
.50	.35	-.15	2.26	1.64	1.68	2.26	3.94	175
	.50	0	80.14	74.52	70.36	74.26	70.62	
.50	.30	-.20	2.16	2.06	1.64	2.70	4.44	99
	.50	0	82.46	74.16	69.98	73.82	70.10	
.60	.40	-.20	2.70	2.76	3.30	2.94	3.04	95
	.60	0	79.70	75.76	75.26	72.52	64.84	
.70	.50	-.20	2.34	2.74	4.54	2.00	1.72	83
	.70	0	79.76	77.38	80.46	70.46	58.14	
.80	.65	-.15	2.22	2.70	7.04	1.46	0.96	112
	.80	0	80.14	77.16	82.96	68.96	54.24	
.90	.80	-.10	2.86	4.86	11.56	2.24	0.88	142
	.90	0	79.70	78.92	86.68	67.48	48.60	

Simulations (N=5000) using SAS RANUNI (SAS, 1989) to generate random variates from a uniform distribution. Sample sizes calculated to show non-inferiority within 10%, 15% or 20% (one sided type I error of 2.5% and type II error of 20%) using approach of Makuch and Simon (1978). To aid interpretation, the magnitude of the Monte Carlo error for the simulation is as follows: 2.5% (SE 0.22%); 80% (SE 0.57%). Further details are provided in the Simulation Note at the end of this Thesis.

As illustrated by Figure 5.4, the PP population produces an unbiased estimate of the odds ratio with the particular formulation of the violator model chosen and the 95% confidence interval is consistently wider when compared with the full population. It followed that the PP population would always be conservative in a regulatory sense and as a consequence the simulation exercises were not repeated for the odds ratio.

Interestingly there is little critical review of the empirical evidence in this area and Ebbutt and Frith (1998) are almost unique in publishing their actual experience of using both PP and ITT populations. They report on the results of 11 asthma trials (sample size ranging from 212 to 421 subjects) that employed an equivalence methodology for a continuous outcome measure – peak expiratory flow rate. These equivalence studies were undertaken in response to the decision to phase out the use of CFCs in aerosols used to deliver inhaled asthma treatment. The ITT populations included all randomised subjects whereas the PP populations excluded subjects who failed to meet the entry criteria (about two thirds of the PVs) or who took proscribed concomitant medication (approximately one third of PV's). The PP populations accounted for between 50% and 88% of the ITT populations (median 75%). Ebbutt and Frith conclude that there was no evidence to suggest a *consistent bias* in either direction when comparing the treatment estimates for the ITT and PP populations. In all cases the width of the 90% confidence interval employed was greater for the associated PP population and given that the residual standard deviations were similar for both populations the authors conclude that *the width of the CI is dominated by the sample sizes in the ITT and PP populations.* Furthermore they suggest that *it is reasonable to base decisions on the ITT analyses provided the PP analyses are supportive.* [Following publication of this Chapter (Garrett, 2003), two statisticians from the FDA (Brittain and Lin, 2005) picked up on the points made above and reviewed 20 anti-infective studies submitted to the agency between 1999 and 2003. They found that in 13 trials the estimate

of the treatment difference was larger for the ITT population compared to the PP population. Furthermore, in 12 cases the ITT confidence interval was wider. Interestingly they reported that *The pattern of anti-conservatism of the PP under these assumptions is consistent with the pattern seen with Garrett's simulation results*. They concluded that *we see no indication that the PP analysis tends to produce a larger absolute treatment effect than the ITT analysis in this setting*, and further speculated that *both analyses may often underestimate the 'pure' treatment difference.*]

5.4.4 Regulatory considerations

The prominent use of the PP population as the primary one for both equivalence and non-inferiority in recent years is most likely a reflection of over interpretation of two related statements in the influential ICH E9 guideline (1998). The first states that the ITT and PP type populations *play different roles* in superiority and in equivalence or non-inferiority trials while the second states that for superiority trials the ITT type population *is used in the primary analysis*. Indeed the European forerunner (CPMP III/3630/92-EN, 1995) to this guideline went so far as to state that *the ITT strategy is insecure* for equivalence studies. However the recent CPMP points to consider document (CPMP/EWP/482/99, 2000) states that ITT and PP populations have *equal importance*. If this latter statement is interpreted as a requirement for both populations to demonstrate equivalence (or non-inferiority), then although the overall type I error rate will be controlled, the type II error will be subjected to inflation. The CPMP's *Points to consider on multiplicity issues in clinical trials* (CPMP/EWP/908/99, 2002) states that when accounting for multiple populations, no adjustment to account for multiplicity is required although no mention is made of the potential impact on power. Interestingly in the same document, for the case of two variables - where statistical significance is required for both - the impact on the type II error is noted and sample size adjustment is recommended. This recommendation

could also be applied to multiple populations although an alternative approach might be to increase power by regarding subject populations (which are based on composite criteria) as if they were like subgroups (which are based on a single criterion such as male subjects). Indeed both are used extensively to assess to the robustness of study conclusions. A step-down procedure could be used to control the type I error whereby the ITT population would be evaluated first as the primary population and if equivalence (or non-inferiority) were shown then the PP population would then be evaluated. This approach is commonly used to evaluate subject subgroups with the aim of controlling the type I error whilst maximising power. Also as described in Section 5.2, adopting an ITT population would facilitate strategies that switch objectives from non-inferiority to superiority (where ITT is required for the primary analysis) or *vice versa*.

Regardless of issues surrounding multiplicity and power, it seems reasonable to conclude that bias is an issue for both PP type and ITT populations. The nature of the bias will depend upon the pattern of non-adherence in the first case and the pattern of missingness and the impact of including non-adherence in the second. However the bias may be in either direction and the PP population is not necessarily anti-conservative for non-inferiority. In fact, one could do worse than employ the eligible population that represents an unbiased (in terms of the randomisation) yet refined population. Fundamentally, however, the key focus of equivalence and non-inferiority studies should be to generate quality data such that data irregularities do not compromise the study conclusions from ITT type analyses. This is a point that is re-iterated by many authors (ICH E9, 1998; CPMP/EWP/482/99, 2000; Lewis and Machin, 1995).

It is worth noting that for the formulation presented earlier, both θ_1 and θ_0 were dependent upon response but independent of treatment. If different values are specified for each treatment then it is simple to create all nature of extreme situations.

A final observation refers to equivalence and non-inferiority studies that incorporate a placebo arm. It is not uncommon to find protocols that plan to use the ITT population to demonstrate study competence (that is, for the comparison of test treatment with placebo) whilst adopting the PP population for the evaluation of equivalence or non-inferiority (that is, for the comparison of test treatment with reference). This is an absurd approach that no logical argument supports. The fact that a trial is competent for the ITT population does not necessarily imply that it is competent for the PP population. Furthermore, if competence has been demonstrated for the ITT population then why not use this to establish equivalence or non-inferiority?

5.5 COVARIATE ADJUSTMENT

5.5.1 The logistic model

Without doubt, logistic regression is the most commonly employed technique to perform covariate adjustment for binary outcome measures in clinical trials. Indeed it is not uncommon to find protocols describing supportive adjusted analyses in the form of logistic regression even in cases where the primary analysis is specified in terms of the difference in proportions. This practice of inconsistent model specification is easily hidden in a superiority framework where hypothesis testing prevails but is exposed when equivalence or non-inferiority margins are necessary. Although stratified analyses do exist for the difference in proportions these can become somewhat unwieldy for more than one or two factors (Smith *et al*, 1998) and suffer from the restrictions imposed by the observed reference proportions described earlier in Section 5.3.

For logistic regression, it is well documented - although not widely known - that if a factor exists which independently affects outcome, then omitting this factor from the model leads to underestimation of a non-unity treatment difference measured on the odds ratio scale (Gail, 1986). As reported in Chapter Three (Section 3.6), this underestimation occurs when the treatment groups are balanced for this factor and as such the observation is applicable to randomised studies. In all cases, the odds ratio shrinks towards unity and the larger the magnitude of the factor effect in relation to a constant treatment effect the more severe the underestimation.

By illustration, take a logistic model of the form, $\log(\pi/1-\pi) = \alpha + \tau_i + \phi_j$, where τ_i indicates the effect of test or reference treatment (i = t, r), ϕ_j indicates the effect of a two-level factor (j = 1, 2) and the odds for each combination are given by $\lambda_{ij} = \pi_{ij}/(1 - \pi_{ij})$. If it is assumed that the number of subjects for each treatment by factor combination is identical, then the unconditional treatment odds ratio (ψ^*_T) – independent of n - can be derived from the individual odds (λ_{ij}) as shown in Chapter Three (Appendix A). That is,

$$\psi^*_T = (\lambda_{t1} + \lambda_{t2} + 2\lambda_{t1}\lambda_{t2})(2 + \lambda_{r1} + \lambda_{r2}) / (\lambda_{r1} + \lambda_{r2} + 2\lambda_{r1}\lambda_{r2})(2 + \lambda_{t1} + \lambda_{t2})$$

Table 5.II shows some examples of underestimation for various combinations of treatment (ψ_T) and factor (ψ_F) effect using a series of reference odds (λ_{r2}). For any particular combination (i.e. row), the extent of the underestimation diminishes as the treatment odds ratio approaches either the relative risk (π_t/π_r) or the inverted reverse relative risk ($1-\pi_r)/(1-\pi_t$) and due to the symmetric nature of the odds ratio this occurs as the odds approach either zero or infinity. (For example, when ψ_T is 0.5 and ψ_F is 4, ψ^*_T is closer to ψ_T for reference odds of 0.125 and 8 than for a reference odds of 1.)

Table 5.II. Unconditional odds ratios (ψ^*_T) produced from a balanced two treatment, two factor design under a range of treatment and factor effects

Treatment OR (ψ_T)	Factor OR (ψ_F)	Reference odds (λ_{r2})						
		0.125	0.25	0.5	1	2	4	8
0.667	1.25	0.667	0.668	0.668	0.668	0.668	0.668	0.667
	1.5	0.668	0.669	0.669	0.670	0.669	0.669	0.668
	2	0.671	0.673	0.674	0.675	0.674	0.672	0.670
	3	0.677	0.682	0.686	0.686	0.682	0.677	0.673
	4	0.684	0.692	0.697	0.696	0.689	0.681	0.675
0.500	1.25	0.500	0.501	0.501	0.501	0.501	0.501	0.501
	1.5	0.501	0.502	0.503	0.503	0.503	0.502	0.502
	2	0.504	0.507	0.509	0.510	0.509	0.507	0.504
	3	0.512	0.518	0.524	0.525	0.521	0.514	0.509
	4	0.520	0.531	0.538	0.538	0.531	0.520	0.512
0.430	1.25	0.430	0.431	0.431	0.431	0.431	0.431	0.431
	1.5	0.431	0.432	0.433	0.434	0.433	0.433	0.432
	2	0.434	0.437	0.439	0.441	0.440	0.437	0.435
	3	0.442	0.448	0.454	0.456	0.453	0.446	0.440
	4	0.450	0.461	0.470	0.471	0.463	0.452	0.443

In epidemiological research the odds ratio is regarded by some as an approximate relative risk since incidence rates are frequently low and denominators are large. However, this is generally not the case in drug development and arguably odds and risks are different measures of outcome and neither should be considered to be an estimate of the other (Hutton, 2000). For the majority of indications, short-term success or failure proportions are frequently in the region 0.5 (λ=1) to 0.8 (λ=4) and clinical trials are typically powered to detect differences of between 0.1 to 0.3 - as such, treatment odds ratios >4 or <0.25 are relatively uncommon. For instance, two treatments with success proportions of 0.8 and 0.5 would generate an odds ratio of 4 that is quite different numerically from the associated relative risk of 1.6 (or the inverted reverse relative risk of 2.5). Interestingly, underestimation of the odds ratio has typically been studied from an epidemiological perspective (Gail, 1986), although it can be seen that the greatest impact is likely to be in the clinical trial setting.

Now, it has also been shown for the logistic model (Robinson and Jewell, 1991; Robinson *et al*, 1998; Lee, 1999), that the exclusion of a prognostic factor leads to an increase in

precision for the estimated treatment difference, rather than a reduction, as one might expect. However, when combined with the underestimation described earlier, there is an overall increase in efficiency (power) from a hypothesis testing perspective (Robinson and Jewell, 1991; Robinson *et al*, 1998) and a strategy of covariate adjustment is therefore justified for superiority studies. (In terms of undertaking a hypothesis test of whether a parameter is equal to zero, the increase in the estimated standard error with the adjusted model is more than counterbalanced by the increase in the maximum likelihood estimate of the parameter and the corresponding Wald statistic will be larger for the adjusted analysis.) However, for non-inferiority, since exclusion of a prognostic factor underestimates a non-unity treatment effect and decreases the associated standard error, both forces could work together to pull the lower confidence limit within a pre-specified margin when compared to the corresponding conditional model. Would this lead to an increased risk of acceptance of non-inferiority when the test treatment was indeed inferior to the reference treatment? A series of simulations has been undertaken to address this point. First, however, it was necessary to create a series of scenarios that were representative of current clinical trial practice.

5.5.2 Supporting simulations

As described in Section 5.3, no accepted margins for the odds ratio exist. As a consequence it was necessary to use those previously recommended for the difference in response percentages and to convert these to plausible odds ratio margins. Non-inferiority margins of 10%, 15% and 20% were considered which translated to odds ratio margins (test:reference) of 0.667, 0.538 and 0.429 respectively for a reference proportion 0.5 which was used throughout. (Note that 0.429 is almost identical to the margin proposed by Tu (1998) and that 0.538 is not dissimilar to both Senn's (2000) margin of 0.55 and the author's proposed margin of 0.5.) A reference proportion of 0.5 is convenient since it enables the response proportions for each treatment by factor combination to be calculated

easily when varying the odds ratio for the factor effect. In each case, the response proportions (π_{r1} and π_{r2}) have been constructed to be symmetric around the overall reference response proportion (π_r) – 0.4 and 0.6, for instance - and can be determined as $[(\psi_F - \psi_F^{1/2})(\psi_F - 1)]$ and $[(1 - \psi_F^{1/2})(\psi_F - 1)]$ respectively by solving the quadratic, $(1 - \psi_F)\,\pi_{r2}^2 - 2\pi_{r2} + 1 = 0$, for a factor odds ratio >1. It is then simple to calculate the corresponding response proportions for the test treatment, (π_{t1}) and (π_{t2}), using the treatment odds ratio (ψ_T) and π_{r1}, π_{r2} respectively.

Table 5.III. Simulation 2: Impact on acceptance/rejection of non-inferiority when excluding a two-level factor from a logistic regression model.

Factor OR	% treatment difference when ψ_T=0.667			Model (N=393)	% non-inferior	
(ψ_F)	Overall	F=1	F=2	per treatment	ψ_T=0.667	ψ_T=1
1	10.0	10.0	10.0	T	2.84	81.86
				F + T	2.72	81.42
2	9.7	10.1	9.4	T	3.04	81.80
				F + T	2.38	79.78
3	9.3	9.8	8.8	T	3.86	81.20
				F + T	2.50	77.46
4	8.9	9.5	8.3	T	5.38	82.66
				F + T	2.56	76.44

Factor OR	% treatment difference when ψ_T=0.538			Model (N=175)	% non-inferior	
(ψ_F)	Overall	F=1	F=2	per treatment	ψ_T=0.538	ψ_T=1
1	15.0	15.0	15.0	T	2.74	82.10
				F + T	2.58	81.94
2	14.6	15.3	13.8	T	2.80	81.14
				F + T	2.20	79.98
3	14.0	15.1	12.9	T	3.62	82.78
				F + T	2.52	79.86
4	13.5	14.8	12.1	T	5.02	82.32
				F + T	2.78	76.52

Factor OR	% treatment difference when ψ_T=0.429			Model (N=99	% non-inferior	
(ψ_F)	Overall	F=1	F=2	per treatment)	ψ_T=0.429	ψ_T=1
1	20.0	20.0	20.0	T	2.42	85.14
				F + T	2.24	83.92
2	19.5	20.8	18.2	T	3.14	84.36
				F + T	2.76	82.48
3	18.8	20.8	16.8	T	3.76	84.72
				F + T	2.44	80.14
4	18.1	20.5	15.7	T	4.54	83.42
				F + T	2.56	78.50

Simulations (N=5000) using SAS RANUNI (SAS, 1989) to generate random variates from a uniform distribution. Sample sizes calculated to show non-inferiority within 10%, 15% and 20% of a reference percentage of 50% (one-sided type I error of 2.5% with type II error of 20%) using approach of Makuch and Simon (1978). Equal number of subjects in each treatment group with random assignment (p=0.5) to each level of factor F. Logistic regression performed using SAS procedure GENMOD (SAS, 1996).). To aid interpretation, the magnitude of the Monte Carlo error for the simulation is as follows: 2.5% (SE 0.22%); 80% (SE 0.57%). Further details are provided in the Simulation Note at the end of this Thesis.

The main finding from Table 5.III is that the type I error is inflated when an influential two-level factor is excluded from the model. When the factor had a large effect ($\psi_F = 4$) then the type I error is approximately doubled. However, the unconditional model provides greater power – the probability of showing non-inferiority when the treatments are truly equivalent. Finally, it should be noted how, with a fixed treatment odds ratio, the overall percentage treatment difference becomes smaller as the factor odds ratio increases while the magnitude of the treatment difference within the subgroups varies reflecting the multiplicative nature of the logistic model.

5.5.3 Regulatory considerations

From a regulatory perspective the specific impact of covariate adjustment on equivalence or non-inferiority determination does not appear to have been considered. Guidance (ICH E9, 1998) of a more general nature exists - such as the inclusion of design features (such as stratification) in the analysis and *a priori* identification of covariates - but specification of the unadjusted analysis as primary appears to be the main recommendation. However, these simulations make a case for the use of conditional logistic models in therapeutic areas with binary endpoints, where there are known factors which impact on outcome and the demonstration of non-inferiority is the primary objective.

5.6 SAMPLE SIZE CONSIDERATIONS

It is well documented that studies designed to demonstrate equivalence or non-inferiority require a greater number of subjects in comparison to superiority studies. The dominant reason for this view appears to be that an acceptable equivalence or non-inferiority margin will always be smaller than the difference that a corresponding superiority trial would be designed to detect. Although this is undoubtedly correct, the point which is missed is that in drug development most test treatments that reach this stage of development are

considered to have at least a small advantage over existing treatments. For instance, a new anti-infective may cover a broader spectrum of pathogens or impact specific pathogens at lower concentrations.

Referring back to Figure 5.1, if the true difference between the treatments is between 0 and $+\delta_2$, then the standard error (SE) required to produce a lower confidence limit (LCL) greater than $-\delta_1$ can be considerably larger than when the true treatment difference is zero. In contrast the SE required to produce a LCL greater than zero – that is, to show superiority – must be much smaller and as such the number of subjects required is larger. Thus for non-inferiority, where only one margin is of interest, the sample size required in practice is likely to be smaller, rather than larger, than that needed for superiority. Interestingly, an earlier version (CPMP/EWP/482/99, 1999) of the CPMP Points to consider document (CPMP/EWP/482/99, 2000) was modified following the consultation process to acknowledge this point. (Note that the author's comments pertaining to this specific point were included as part of the PSI [Statisticians in the Pharmaceutical Industry] response to the CPMP.)

Equivalence is different, however, as both confidence limits must lie within the equivalence margins and if the true treatment difference is between 0 and $+\delta_2$ then a smaller SE is required for the UCL to be less than $+\delta_2$ than for the LCL to be greater than $-\delta_1$. However, as stated earlier, non-inferiority rather than equivalence is usually applicable to the therapeutic setting.

It is standard practice to base the planned sample size on the PP population for equivalence and non-inferiority trials. This is consistent with the current trend to specify the PP population as primary but is also justified if both populations have equal standing

or even if the PP population is secondary since a sufficient number of subjects is required to enable reliable conclusions to be drawn. However, if the ITT and PP populations are to have equal standing (CPMP/EWP/482/99, 2000) and equivalence or non-inferiority is to be concluded *only on the basis of strict adherence to the margins for both populations* then there is potential for an overall increase in the Type II error. Alternatively significance may not have to be achieved in both populations - although inevitably this leads back to the pre-specification of a primary population to avoid inflation of the Type I error.

The final point regarding sample size calculation refers to covariate adjustment. Since the inclusion of a influential factor in the logistic model increases the standard error of the estimated treatment difference (rather than reduces it, as per ANCOVA), this has led some to conclude that the sample size for superiority studies must be adjusted upwards accordingly (Hsieh, 1998). However, as has been shown earlier, this is simply not the case since consideration has to be given to fact that the parameters being estimated in the two model formulations – unconditional and conditional - are actually different and the increase in SE is counterbalanced by an increase in the magnitude of the parameter being estimated. As such, effort to increase the planned sample size to account specifically for covariate adjustment for superiority trials is misplaced. In a similar vein, Whitehead (1993) incorrectly asserted that an increase in the sample size was required for the related proportional odds model. Ironically, however, variance inflation is pertinent to both equivalence and non-inferiority trials. When two treatments are truly equivalent, adjustment for a prognostic factor will inflate the standard error and reduce the probability of showing equivalence or non-inferiority within given margins (that is, power). As such, the sample size will need to be increased to maintain power at the pre-specified level.

5.7 DISCUSSION

The primary aim of this chapter is to challenge conventional thinking in the area of therapeutic equivalence and non-inferiority. As the number of equivalence and non-inferiority trials has grown, the regulatory authorities that govern the approval of pharmaceutical treatments have acted quickly to provide a framework for conducting such studies. However, in many ways their advice remains untested and the implications of their advice in some instances is not fully understood. The intention of this research, however, is not to provide a set of specific solutions to the problems raised, rather some suggestions are made which could lead to a more coherent approach.

One particular approach that could confer advantages for binary outcome measures is the specification of margins in terms of the odds ratio. Sound analyses begin with sound designs but the regulators have until very recently sorely neglected margin specification, which is so fundamental to the equivalence and non-inferiority methodologies. Previous efforts in the specific area of anti-infectives appear to have been misplaced and have led in some cases to data driven criteria. As described, the odds ratio has many desirable properties. In particular it is stable over a wide range of conditions and is easily incorporated into the generalised linear model framework to facilitate covariate adjustment. In contrast, approaches based on the difference in proportions are unwieldy and have many limitations. Unlike the logit transformation where parameter values are in the real plane $(-\infty,+\infty)$ the difference in proportions is bounded in the unit square $(-1,+1)$. This becomes highly restrictive in cases where the reference treatment has response proportions approaching 0 or 1. Furthermore the Normal distribution approximates the binomial poorly in these cases and can lead to improper confidence limits. That being said, many researchers are more comfortable working with proportions and have less of an intuitive feel for parameters or summaries expressed in terms of odds ratios. As such, an

odds ratio based approach might face some resistance on the grounds that it is too obscure. However given that the whole concept of equivalence and non-inferiority margins is regarded by some to be somewhat obscure, the odds ratio may provide an opportunity to provide a more coherent approach to the problem. A non-inferiority margin of 0.5 would provide a useful starting point for further discussion since it ties in well with some original regulatory strategies.

There are known concerns surrounding the choice of analysis populations for equivalence and non-inferiority although as has been shown it is not clear that PP type populations necessarily provide a solution. The perceived conservative nature of the PP population appears to be much more a reflection of reduced subject numbers than the presence of bias while bias can be in either direction depending on the pattern of violations. It is the author's opinion that PP type populations should be used with caution – much more so than for ITT type populations. If a primary population is to be selected for equivalence or non-inferiority then one could do worse than choose the eligible population since this is unbiased in terms of the randomisation but provides a refined set of subjects who should be sensitive to demonstrating real treatment differences.

At this stage it is unclear what the practical implications will be of implementing a regulator's rule of equal importance for the ITT and PP populations. What is clear is that when either switching objectives – from non-inferiority to superiority or *vice versa* - or when using a placebo group to establish assay sensitivity, the only logical approach is to use the same primary population throughout the sequence of analyses. Given the regulatory authorities' preference for ITT in the case of superiority, it is difficult to see how PP type populations could fulfil this role. If there is consensus, then it is that quality

data are of prime importance if reliable conclusions are to be drawn from equivalence and non-inferiority studies.

It has been shown that modest inflation of the type I error rate could occur when an important factor is omitted from a logistic regression model. Inflation of the type I error will also occur with the hazard ratio and with the odds ratio from the proportional odds model, although this has not been quantified in this research thesis. Of these two models, this observation is likely to be more relevant to the hazard ratio since increasingly this parameter is being used to set margins for non-inferiority trials with long term outcomes – for example, mortality or cardiovascular endpoints in diabetes. As such, conditional analyses should be recommended in cases were a study is stratified and/or a factor is known to have a large impact on outcome. This recommendation holds true regardless of whether the hypothesis to be tested relates to equivalence, non-inferiority or superiority, and again it is important to have consistency if there is a plan to allow the switching of objectives between non-inferiority and superiority within a trial.

Interestingly the ICH E9 (1998) statistics guideline makes the general claim that covariate adjustment leads to an increase in precision. In fact for a whole class of models commonly used in clinical research - including the logistic, proportional hazards (Ford, 1995) and proportional odds models – a reduction in the standard error of the estimate does not occur with covariate adjustment. Although one could argue that in general there is an increase in relative precision (in terms of the Wald statistic), to avoid potential confusion it is recommended that the guideline is modified to state that covariate adjustment leads to an increase in efficiency rather than precision.

Finally, although it has been shown that an increase in sample size is not needed for superiority studies when covariate adjustment is planned, an increase is likely for equivalence or non-inferiority studies when the treatments are truly equivalent to maintain power at the planned level.

In the following chapter (Chapter Six), attention turns to the unique challenges faced when undertaking drug development in special populations - in this case, children.

CHAPTER SIX: THE THERAPEUTIC ORPHANS.

Jennifer Eccles

Had terrible freckles

And the boys all called her names

But she changed with medicinal compound

And now she joins in all their games.

6.1 INTRODUCTION

Since children represent around one half of the world's population of six billion persons, it is perhaps surprising that, in relation to drug development, paediatrics is viewed as a special interest population. However, as long ago as 1968, Dr Harry Shirkey coined the phrase *therapeutic orphans* to highlight the fact that the labels of pharmaceutical treatments mostly discouraged their use in children (Kearns, 1996; Coté *et al*, 1995). This situation had arisen not because drugs were necessarily unsafe in children – but rather as an indirect consequence of a dearth of clinical trial data in this area. Although much of the regulation that had been introduced to control the marketing of drugs was in response to tragedies involving the treatment of children, in practice such regulation had failed to deliver safe and effective treatments to the very subjects it was intended to protect. Despite being noble in aim, the regulation had in effect discouraged paediatric drug development and pharmaceutical companies, finding little financial incentive to conduct paediatric studies, side-stepped the issue for fear of liability and allegations of malpractice. To this day, the viewpoint of pharmaceutical executives is all too familiar – that is, in many instances, paediatric studies are challenging to perform, difficult to justify from an ethical standpoint, and due to the limited market size, not financially rewarding enough (USA Today, 20 December 2000).

Nonetheless children become ill and require treatment and the lack of specific labelling information does not mean that children do not receive drug treatment. Rather they receive what is described as off-label treatment. In the context of paediatric medicine, this phrase refers to the current world-wide practice of prescribing drugs to children for a different indication, or outside of the age range, for which regulatory approval has been granted; an act that in effect creates uncontrolled and undocumented experimentation in children. For instance, in 2001 the General Accounting Office (GAO) in the US stated that around 75% of all marketed drugs did not include labelling information for children (GAO, 2001). Furthermore, the younger the children, the more likely that pertinent information was absent from the label (Roberts *et al*, 2003). In the UK, Sweden, Germany, Italy and the Netherlands it was observed that more than two-thirds of the children who were admitted to paediatric units received unlicensed or off-label medication. Furthermore, for many of the drugs prescribed no paediatric formulation existed and therefore prior to administration some modification to the manufactured product was required (Pharmaceutical Executive, 1 February 2000). In the new age of global bio-terrorism it will most likely be children that will be the most vulnerable due to their size, developing immune systems and higher respiratory rates (such that contaminants are breathed quicker). However the most effective treatment for anthrax (ciprofloxacin) is not recommended for children and there is in general a lack of research in the area of antidotes for children (Associated Press Newswires, 17 October 2001). This is not to say that off-label prescribing is necessarily considered malpractice - indeed the American Academy of Pediatrics has actually highlighted that the failure to use off-label drugs where appropriate could instead represent negligence on the part of the physician (Banner, 2002). Indeed Budetti (2003) describes off-label prescribing *as the cornerstone of pediatric medical therapeutics* – his argument based on the fact that restricting prescribing to approved drugs would deny children access to many modern medicines and

provide only a *relatively modest pharmacopoeia.* However, despite such encouragement, it is not the case that the use of off-label treatments has been without its problems and the paucity of paediatric data has led to some notable medical disasters. For instance, verapamil which had been used safely to treat adults for heart disease was given to infants during the 1980s to treat irregular heartbeat and unexpectedly caused cardiac arrest, while in the1990's, a local anaqsthetic unexpectedly caused seizures in children aged under 12 years due to overdosing. (Although not taken directly by children, thalidomide was banned in 1962 after causing birth defects in nearly 10,000 children as a result of it being given as a sedative to pregnant women to treat morning sickness.)

There are, however, several therapeutic areas that represent notable exceptions where substantial paediatric drug development has been conducted. These include the areas of vaccination, bacterial infections, AIDS and oncology. It has even been reported that around 70% of children in the US with cancer are enrolled in clinical trials (MJ Kupst speaking on a panel at the 2001 American Psychological Association meeting, reported in The Arizona Republic, 5 September 2001). Typically it has been the more specialised drugs that have been tested in children and indeed it was the development of drugs to treat AIDS that helped bring the paediatric drug testing issue to the fore in the 1990's. However it is also important to note that it is considered unnecessary to test in children drugs aimed at adult only indications such Alzheimer's disease, Parkinson's disease and breast, prostrate and lung cancers. It is also worth noting that in the broader context of child health it is estimated that 86% of all children are born in less developed countries, where 97% of all child death occurs (Schaller, 2000). (For instance, each year one million children under the age of five years die from malaria.) Thus although the children of the developing world share the same problem of off-label prescribing as those children in

developed countries, the more acute problem is not so much about off-label use of drug treatments, but with access to any drugs at all (Milne, 2000a).

The aim of this Chapter is to illustrate some of the unique challenges that must be overcome in order to provide safe and effective drug treatment for the illnesses of childhood. Firstly, the reasons why children represent a special population and why drug development in children differs from drug development in adults will be presented. Secondly, the evolution of paediatric drug development will be discussed with particular focus on the initiatives introduced in the US over recent years. Thirdly, current regulatory thinking and guidance in the area of paediatric drug development will be reviewed. Finally, some challenges related to trial design will be discussed – in particular the choice of control group and the long-term follow up of subjects.

6.2 GENERAL CONSIDERATIONS

Dianne Murphy of the FDA has captured the main challenge of developing drug treatments for children and has illustrated why children represent a special population. She notes how children have historically been incorrectly treated like smaller versions of adults whereas in reality they represent *unfinished products*. Furthermore, she explains how children are continually developing – *different things turn on and off, different enzymes, different receptors become inactive or active, and because of that, their susceptibility and responses are different* (Chicago Tribune, 1 December 2000). These descriptions serve to illustrate the unpredictable nature of the administrating drugs to children since one has to consider not only the large variability associated with what the child's body does to the drug (the pharmacokinetics [PK]) but also what the drug and resulting metabolites do to the child's body (pharmacodynamics [PD]). As such, one also

has to consider what the impact on the finished product will be when it matures in years to come.

Infants, and in particular premature babies, exhibit rapid physiological change that is particularly challenging for safe and effective drug treatment. The adult body composition of 60% water is not achieved until a child reaches one or two years of age and will initially be more in the region of 80% (Schaller, 2000). Around the same time, renal systems mature (Spielberg, 1996). Renal elimination is particularly important for water-soluble drugs and metabolites. For instance the functional capacity of renal tubular secretion, that has a large impact on renal clearance, increases two-fold during the first week following birth and increases ten-fold over 12 months. The glomerular filtration rate typically reaches adult levels after 6 months - although can be highly variable up to this point (Reed, 1996). Relative gastro-intestinal surface area is larger in younger children although this does not necessarily lead to greater absorption of oral drug treatments. In this case, absorption is governed by gastric emptying, transit time and intestinal motility which in turn are influenced by dietary changes – such as the introduction of solid food. Furthermore the interaction between the pH of the drug in question and local gastric pH (which is around 7 at birth and decreases gradually to the adult level over the next two years) impacts the absorption of the specific drug concerned (McRorie, 1996; Reed, 1996). Elimination of most drugs from the body occurs as a result of biotransformation to water-soluble metabolites, and in this context the cytochromes P450 are the most important hepatic enzymes. However, hepatic expression of P450 enzymes is not monotonic with age, since although P450 activity is limited in newborns, it actually exceeds adult capacity in young children and only begins to decline at puberty. Furthermore inter-individual variability in the relative expression of specific P450 enzymes means that clinical response

can be markedly different since different P450's produce different metabolites from the same drug (Leeder, 1996).

Formulations should be developed to ensure the safety and consistency of administration (Roberts, 2002) but the route of administration in children also requires careful consideration – especially in the very young. Intra-muscular injections may cause tissue injury while transdermal applications can lead to systemic toxicity from increased skin permeability – although three weeks after birth a full barrier will exist. From a practical perspective, premature infants can tolerate only small volumes of non-nutritive fluids (McRorie, 1996). A further consideration for children is the impact of preservatives that are contained in pharmaceuticals products. In particular multiple doses with the same preservative can lead to life threatening toxicity as illustrated by the case of the preservative benzyl alcohol in sodium chloride flush solutions. The oxidation and conjugation steps in neonates is immature and in the early 1980's this led to 16 deaths as a result of benzyl alcohol and benzoic acid accumulation (McRorie, 1996).

Of course drug treatment in babies is not only limited to direct drug exposure. Babies may receive drugs indirectly via their mother either when in the womb or through breast milk. Indeed it was through the maternal route that thalidomide impacted limb growth in the womb while NSAIDs have been linked with persistent pulmonary hypertension (PPHN) in newborns (Alano *et al*, 2001).

As children develop, due consideration needs to be given not only to the impact of physical growth and sexual maturation on drug disposition, but also the effect of drug disposition on these aspects of child development - including mental development. As children approach adulthood, self-medication also becomes a feature as independence

from parental control increases. The impacts of alcohol, tobacco and illicit drugs also need to be considered, as do abnormal sleeping or eating habits. From a drug development perspective, this movement towards active participation is important as children come to understand the risks and benefits of being enrolled in clinical trials and are able to provide informed assent - even though parental consent is also required. In contrast to small children where caregiver and parental assessment of response is dominant, older children begin to be able to provide their own assessments of efficacy and toxicity.

Historically the standard approach to the treatment of children with adult drugs has been rather crude. The adult dose has generally been adjusted for weight in younger children while tablets have been crushed and mixed with food to facilitate ingestion. (For instance, the starting dose for Phase I paediatric oncology is typically 80% of the maximum adult dose with dose escalation in increments of 20 to 30% (CPMP Addendum on paediatric oncology, 2003).) However, as illustrated, there are many other factors that influence how drugs are distributed through the body – including ingestion with food itself. As a consequence, inadequate dosing information and inappropriate formulations can limit efficacy (due to under dosing) or increase the risk of toxicity (due to over dosing) in children. (For instance, in the treatment of preadolescent girls for obsessive-compulsive disorder it was found that the dose of Luvox being used was at least twice as high as necessary (D Murphy of the FDA reported in The Wall Street Journal, 5 February 2001).) In this respect, the weaknesses of the old dosing habits serve to highlight two key steps for successful paediatric drug development – that is, appropriate formulation development and effective PK investigation.

Age appropriate formulations are required for marketed products and also for experimental treatments to ensure accurate, reliable and suitable drug administration. (Note however

that weight based dosing is often preferable to age based dosing for younger children (under 6 months, say) where age should only be used if weight is unknown (Tan sheet, 3 September 2001).) Examples of paediatric formulations include chewable tablets and in particular liquids for children under 5 years of age. For instance to formulate an oral product for paediatric use, one typically dissolves or suspends a drug in a simple syrup or sorbitol – although some children may have fructose intolerance so care needs to be exercised. Needless to say when paediatric formulations are not available, physicians have to compromise. For instance, in a trial to treat babies for a rare but potentially fatal lung disorder, investigators had no other option when using Viagra experimentally than to crush a pill and administer part of the contents through a feeding tube (Atz and Wesel, 1999, as reported by Associated Press Newswires, 3 August 2001). However tablet destruction can have a marked effect on a drug's bioavailability. Similarly, a US survey reported that that more than 90% of parents have had to mix drug treatments with food to get their child to swallow and, of these, most were unaware that such a process could impact efficacy (PR Newswire, 7 April 2003). According to Roberts (2002), paediatric formulations should be *palatable and permit easy dose titration*. There are also practical aspects such as the dosing schedule for children of school age. In this case there are clear advantages in developing once daily regimens of drugs that ease the burden of school nurses whilst reducing both the social stigma of taking drug treatments and the risk of drug misuse. However the development of once daily regimes is not without challenges either. For instance, a regulatory safety concern was raised for a once daily formulation of a treatment for attention deficit hyperactivity disorder (ADHD) with regard to tablet size and the risk of intestinal obstruction. This finally led to a contraindication relating to pre-existing severe gastrointestinal narrowing (Pharmaceutical Approvals Monthly, 1 January 2001). For intravenous formulations the challenge is to provide an age appropriate volume. For example, in seriously ill neonates with fluid restrictions, a significant

proportion of their daily fluid intake can be made up of drug infusions. As such, the aim should be to restrict the total volume of medication to <10% of the average daily fluid intake while post-infusion phlebitis can be avoided by medication dilution.

An important step in development of any paediatric formulation is the comparison in adult subjects of this formulation with the standard adult formulation with regard to the PK profile. This step then leads to the selection of an appropriate age or weight specific paediatric dose for further PK study in the paediatric population. In this respect the standard approach is to mimic the adult PK profile in children through dose adjustment and to extrapolate the clinical efficacy from the adult clinical studies. Indeed this philosophy is adopted throughout the whole clinical development process as a means of determining the appropriate dose in various subgroups including different ethnic groups. ICH E5 *Note for guidance on ethnic factors in the acceptability of foreign clinical data* (1998) describes the process for using foreign data to support drug approval in another geographical region. A so-called bridging package is prepared that contains PK, preliminary PD and dose response data plus a bridging study actually conducted in the new region that allows extrapolation of the original foreign clinical data to the new region. In some cases, the bridging study may simply be a PK study conducted in the new region in the requisite ethnic population. The approach is not without some risks in the paediatric population since the diseases of children and adults can differ and even when diseases are the same, the observed adverse event profile in adults may not necessarily predict the profile in children (Roberts, 2002). Further considerations in determining the paediatric PK profile is consideration of the impact of formula milk since, before a child moves on to solid foods, medication is often mixed with infant formula and the stability of the drug needs to be demonstrated.

6.3 THE EVOLUTION OF PAEDIATRIC DRUG DEVELOPMENT

Paediatric drug development is a recent phenomenon which is most advanced in the US where the FDA has driven a whole series of paediatric initiatives over the past 20 years or so. In 1979, the FDA published its first paediatric labelling regulations to correct the problem that many of the drugs that included a paediatric disclaimer - such as safety and effectiveness in children have not been established - were actually used in children. The regulation required drug companies to conduct adequate and well-controlled trials in order to produce clinical data such that appropriate information could be incorporated into the drug label. However the regulation was generally not regarded as a great success and the expected increase in activity did not materialise.

Fifteen years later the FDA tried again and issued a new Pediatric Rule in December 1994 that softened the previous requirements somewhat. The rule allowed the extrapolation of data from adequate and well-controlled adult trials to children in cases where the therapeutic indication was similar and the drug was expected to behave in the same way. These data would then be supplemented by PK and safety data from the paediatric population. Alongside the rule, a Pediatric Plan was published to direct attention to the whole drug development process in relation to paediatrics. Interestingly the rule did not actually require companies to conduct paediatric studies (Roberts and Maldonado, 1996).

However it was the FDA's Modernization Act (FDAMA) of 1997 that provided the real impetus for the development of drugs in children since it gave pharmaceutical companies a direct financial incentive to conduct paediatric studies. The provision provided a further 6 months of marketing exclusivity for companies that voluntarily undertook studies on approved drugs that the FDA specifically targeted for paediatric development. (Such marketing exclusivity essentially provides protection against the generic competition that

erodes profit when patents expire.) For their part, the FDA was required to publish a list of approved drugs that were currently used off-label in children and for which paediatric data were needed (Roberts *et al*, 2003). A drug had to meet one of three criteria to be included on the list - it had to represent a *significant improvement* compared with drugs currently labelled for paediatric use; it had to be *widely used* in children (at least 50,000 projected uses per year); and it had to be in a class or indication for which additional *options* were needed for children (GAO, 2001). The aim of the list was to prioritise the drugs for paediatric development and in practice was produced with the help of other learned bodies - such as the American Academy of Pediatrics. (It is intended that in future, the FDA will augment such advice with data on actual off-label prescribing habits provided by commercial databases (Pink Sheet, 22 October 2001).) In fact, most of the FDA's subsequent requests for paediatric data have pertained to these lists – produced annually - and the 1997 provision was later renewed in 2002 through The Best Pharmaceuticals for Children Act.

A number of reviews assessing the impact of the FDAMA in effecting drug development in children have been conducted. A GAO report issued in 2001 noted a substantial increase in both the number of drugs studied in children and the range of therapeutic classes covered since enactment, and also noted marked growth in the infrastructure required to conduct paediatric studies. As of April 2001, 28 drugs had been granted marketing exclusivity extensions while the labels of 18 drugs had been updated with new and useful information. The GAO did observe however that there remained a requirement to tackle those commonly used drugs where the patent had expired (six of the top ten most commonly prescribed treatment were found to be off-patent in 1994) and where the financial incentive to undertake paediatric development did not exist.

A year on, Roberts *et al* (2003) detailed the FDA perspective of the impact of the FDAMA - that is, from July 1998 to April 2002. During this period, 53 drugs had been granted marketing exclusivity extensions while 33 drugs had had their labels updated with paediatric information. (Indeed the pace of progress is indicated by the fact that, as of January 2003, 16 more drugs have had label modifications such that the total is now 49.). In total, over 50,000 subjects have been enrolled in studies covering a wide range of conditions including, amongst others, allergies, anxiety, diabetes, epilepsy, gastro-oesophageal reflux, HIV, hypertension and rheumatoid arthritis. Of note were data from over 40,000 subjects that lowered the threshold for Ibuprofen in the treatment of fever from 2 years to 6 months. Furthermore, specific paediatric formulations had been developed for five drugs. Of the studies performed, 23% were safety only investigations, 34% were classified as safety plus PK/PD, while the remaining 43% were classified as safety plus clinical efficacy. Interestingly, for the 33 drugs with updated labels the tendency has been for the new data to pertain to older (>3 years) rather than younger children although all but one of the label changes failed to lower the age threshold for which information on the label was provided. In response to this it has been suggested that an additional 3 months of exclusivity (in addition to the original 6 months) should be available to pharmaceutical companies for conducting studies in neonates (Generic line, 7 September 2001).

The creation of the mandatory Pediatric Rule in 1998 was, however, a less successful development. The intention was that this would complement the voluntary FDAMA since it required drug companies to conduct paediatric studies for experimental drugs currently being investigated in adults where the benefit or use in the same indication was likely in children. In October 2002, the American courts actually blocked the FDA from enforcing such a rule and in response The Best Pharmaceuticals for Children Act of 2002 was

established. This introduced a procedure to investigate drugs that were off-patent and also set up a research fund to conduct such studies. However in late 2003, the Pediatric Research Equity Act was signed which now enables the FDA to require pharmaceutical companies to test both new and existing medications on children.

The FDA's Pediatric Advisory Committee has also been influential in addressing specific issues in paediatric drug development including: the recommendation to study patients - who may benefit from participation in a trial - rather than healthy volunteers (1999); a preference for the enrolment of children who are able to give assent (2000); the acceptability of placebo controlled trials (2000); and the protection of vulnerable paediatric populations (2001). Also the impact of the WHO over the past two decades should not be forgotten. The fact that current treatments were either not approved for children or in some cases unaffordable, led the WHO to develop the essential drugs program as early as 1981. In this respect, essential drugs are those *most necessary for the health needs of the population and which should be made available in regular supply at the lowest cost* (Milne, 2000b), and it is encouraging that in the first year of the FDAMA program, 10% of FDA's requests for paediatric studies were for drugs on the list. However it is clear that the US has been the leader in paediatric drug development and undoubtedly its series of initiatives will have a major impact on child health globally, as paediatric data are shared across regions.

In the following section the latest guidelines governing paediatric drug development will be reviewed.

6.4 REGULATORY CONSIDERATIONS

Primary world-wide guidance is provided by ICH E11 *Note for guidance on clinical investigation of medicinal products in the paediatric population* (CPMP/ICH/2722/99) which was issued in 2000 with the goal of encouraging and facilitating paediatric drug development. The guideline emphasises the timeliness of the initiation of the paediatric drug development program in relation to several factors that define the risk/benefit ratio - for example, the seriousness of the condition. Generally, unless the condition is serious or life threatening, or the indication exclusive to the paediatric population, then a paediatric program should not normally commence until safety and preliminary efficacy have been demonstrated in adults. However life-threatening conditions where treatment options are limited, or indications specific to paediatrics, warrant early initiation following initial safety studies in adults.

In all cases, PK studies are seen to play an important early role - specifically in the areas of formulation development and dose determination. Formulation development is viewed as crucial in the ultimate determination of accurate dosing in paediatrics and is key to ensuring compliance. As a first step, studies comparing the paediatric formulation with the adult formulation should be conducted in adults. Then, once an acceptable formulation is produced, the general approach for indications common to both adults and children, is to determine an appropriate paediatric dose by attempting to mirror the adult PK profile (that has been shown to be therapeutically effective) in older children. In this respect the adult dose is usually adjusted for body weight (mg/kg) or, less commonly, body surface area (mg/m^2) and the reasoning is that similar blood levels of the drug in children compared with adults will accordingly produce corresponding efficacy and safety. Indeed the same rationale is then used to extrapolate from older children to younger children further along the development program. It is important to note that unlike adult studies, which are

usually conducted in healthy volunteers, paediatric PK studies are performed on patients. Now, according to the guideline, for indications where both the disease process and the outcome are similar in children and adults, satisfactory PK data supplemented with acceptable safety data should be regarded as sufficient to extend drug approval to the age range of children covered by the paediatric data. However in cases where equivalent drug concentrations are not expected to produce similar outcomes in children, the guideline requires additional PD studies to be conducted. (For example, the investigation of gastric pH and the duration of acid suppression for drugs used to treat some gastrointestinal diseases.) Similarly PD studies would be required for topical formulations. Finally, for novel paediatric indications or where the disease course or outcome is expected to be different in children compared with adults, adequate and well and controlled clinical trials of clinical outcome and safety are expected before paediatric approval would be considered. Regarding efficacy studies, the guideline highlights the potential need *to develop, validate and employ different endpoints for specific age and development groups.* In particular the measurement of subjective symptoms, such as pain, gets specific mention. However it is important to note that ICH E11 is intended for use in conjunction with many other ICH guidelines. As such many of the basic design, conduct and reporting guidance detailed in documents such as ICH E9 and ICH E10 apply equally to the paediatric population and are not repeated in ICH E11. Safety data are required for all scenarios with the acknowledgement that unique adverse events may occur in the paediatric population or in subgroups of the population given the diversity of maturation in children. In particular the long-term follow-up of subjects receives special mention in relation to delayed drug effects and the impact of chronic drug treatment on *skeletal, behavioural, cognitive, sexual and immune maturation and development.*

With regard to age, ICH E11 suggests five categories or strata whilst acknowledging that they are somewhat arbitrary and that for a specific drug, consideration needs to be given to developmental biology and pharmacology. The guideline also acknowledges that children could actually move categories during a study. The categories specified are: pre-term new-born infants; term new-born infants [neonates] (0 to 27 days); infants and toddlers (28 days to 23 months); children (2 to 11 years); and adolescents (12 to 16-18 years, dependent on region). Some specific challenges noted include: the extrapolation from any other category to the pre-term new-borns; the unpredictability of oral absorption and the immaturity but rapid change of hepatic and renal systems (0-27 days); rapid CNS maturation, immune system development and total body growth (28 days to 23 months); renal clearance exceeding adult levels, psychomotor development and puberty in girls (2-11 years); and sexual maturation and pregnancy testing in girls (11 years upwards).

Ethics is the final area covered by the ICH E11 guideline. It specifies that *special measures are needed to protect the rights of participants and to shield them form undue risk* and that ethics committees should therefore be knowledgeable of paediatric research. Furthermore those conducting the research should be skilled in performing procedures in children and be properly trained and experienced. It is noted that children cannot provide informed consent to participate in a clinical trial and parents or a legal guardian must therefore assume that responsibility – although child assent is encouraged where appropriate and participants should be made fully aware of their rights. It is also expressed that if the desired information can be obtained in a less vulnerable population that is able to provide consent then this population should be used instead. Minimising risk and distress is important and researchers are expected to be proactive in this area. Study procedures should be designed for children and in particular minimally or non-invasive procedures are encouraged.

In Europe, a *Notes for Guidance* document entitled *Clinical investigation of medicinal products in children* (CPMP/EWP/462/95, 1997) was issued as early as 1997 although this has effectively been superseded by ICH E11 - since the former was incorporated into the latter. However there is now a growing recognition in Europe of the need for incentives to encourage paediatric drug development together with the need for more specific guidance to complement ICH E11. Regarding the direct encouragement of paediatric drug development in Europe, a consultation paper was released in February 2002 entitled *Better Medicines for Children.* This discussed the possibility of an exclusivity programme based on the US model but with possibly 12 months provision. Furthermore, in relation to further guidance, the CPMP has issued two concept papers in 2002 stating their intention to generate specific guidelines on the evaluation of the pharmacokinetics of drugs used in the paediatric population (CPMP/EWP/968/02, 2002) and the conduct of pharmacovigilance activities (CPMP/PhVWP/4838/02, 2002).

The underlying themes of the first concept paper were the use of PK studies to support formulation development, age specific dose recommendations and the extrapolation of adult efficacy data to children. The requirement for more information on design, analysis and interpretation was noted and issues likely to be addressed in the subsequent guideline included: identification of the important PK parameters in children; age classification and stratification; and the evaluation of PK in neonates.

The second concept paper relating to pharmacovigilance is unusually detailed. It highlights that compared with adults, the clinical safety database for children is likely to be sparser at the time of regulatory approval - particular in rare indications. Furthermore long-term data are required to detect delayed toxic effects or those due to chronic

medication use - particularly in relation to child development; both of these factors underline the need for enhanced phramacovigilance in children. The proposed approach is to identify the limitations of the safety database at the time of registration and the potential risks of drug approval with the aim of recommending post-approval data collection mechanisms to minimise such risks. The two key elements are signal detection and signal evaluation with the spontaneous reporting of adverse events remaining the most important data source. The concept paper highlights that children may not effectively express symptoms and that parents represent an additional intermediate step in the reporting process such that some sort of facilitation needs to be devised. The role of Periodic Safety Update Reports (PSUR) in relation to children will also be reviewed, as will be the systematic search of the paediatric literature. Post-approval studies are classified as safety demonstration (large studies); new safety issue detection; and known safety issue evaluation. Advice will be developed in relation to when each category of study should be conducted and the methodologies to be used to detect the longer-term effects of treatment. Interestingly the concept paper will give due consideration to paediatric pharmacovigilance activities when treatments are used off-label – perhaps even providing tacit approval for this practice. Vaccines will be given special consideration due to the sensitivity of administering drugs widely to large and healthy populations. The concept paper explains that *even rare suspected adverse drug reactions following vaccination* require concentrated investigation and that follow-up must be addressed in relation to delayed effects. It highlights that the size of the clinical trial programme is driven from the efficacy perspective and that the requirement for effective post-approval pharmacovigilance is an inevitable consequence if safety is to be demonstrated long-term.

Also in Europe, the idea of a Paediatric Addendum to therapeutic specific guidelines has recently been introduced. For example, in 2003 the CPMP issued an *Addendum on*

paediatric oncology (CPMP/EWP/569/02, 2003) to the *Note for guidance on evaluation of anticancer medicinal products in man.*

Further regulatory developments in the US include a draft FDA guideline issued in July 2003 directed towards paediatric medical devices (Device and Diagnostic Letter, 28 July 2003). Unique paediatric challenges identified in this document include the longevity of devices and the impact of such devices on growth and development. Proposed age strata of interest where given as: 0 - 1 month; 1 month - 1 year; 2 years - 12 years; and 12 years to 21 years. Although almost identical to the strata specified in ICH E11, it is interesting to note the higher upper limit of 21 years for devices compared with the 16-18 years ICH E11 limit (dependent on region) for drugs. Other developments are likely to include a FDA guideline aimed at directing pharmaceutical companies to assess the amount of drug and metabolite in breast milk in order to provide dosing recommendations for women who plan to breast-feed. In this respect, mothers still producing milk after having weaned their babies have been identified as being especially useful for collecting data (The Pink Sheet, 30 July 2001).

6.5 SPECIFIC DESIGN ISSUES: CONTROLS AND FOLLOW-UP

In the modern era, the strict application of ethical principles to the participation of humans in clinical research was a direct result of the atrocities committed during the Second World War. In the early post war years (1946 – 1949) the so-called Nuremberg Code (1949) was introduced as a positive outcome of the successful trials of war criminals in Nuremberg, Germany. The Nuremberg Code established that, within reasonably well-defined bounds, clinical experimentation was indeed ethical and yielded benefits to society as a whole. However certain basic principles needed to be observed to satisfy moral, ethical and legal concepts. Voluntary consent of participants was identified as *absolutely essential* and is

the first element of the ten-point code. As the same time the participant should have *sufficient knowledge and comprehension of the elements of the subject matter* so as to make an informed decision. Yet the Nuremberg Code was in effect superseded in 1964 by the introduction of the World Medical Association Declaration of Helsinki and to this day, this 32-point declaration remains central to the conduct of ethical clinical research – albeit in amended form (WMA, 2000).

According to David Wendler of the National Institute of Health, the abuses and atrocities of the Nuremberg trials era had a lasting effect on the research psyche, and led to increased sensitivities world-wide regarding informed consent and the participation of vulnerable populations in clinical research – particularly children. However in his view, *the pendulum is swinging back*, with growing awareness of the need to manage not only the risks of clinical research but also the potential benefits (USA Today, 20 December 2000).

In this sub-section, two specific issues surrounding participation in paediatric clinical trials will be discussed in relation to informed consent and risk management. The first is the choice of control group while the second is the long-term follow-up of children.

6.5.1 Control groups

The choice of the control group raises unique challenges in paediatric drug development. However to understand these challenges it is important to review first the general considerations faced when determining the acceptability of specific types of control in drug development.

The Declaration of Helsinki (DH) is central to the conduct of clinical trials and in relation to control groups the recent Edinburgh revision (WMA, 2000) states the following. *The benefits, risks, burdens and effectiveness of a new method should be tested against those of the best current prophylactic, diagnostic, and therapeutic methods. This does not exclude the use of placebo, or no treatment, in studies where no proven prophylactic, diagnostic, and therapeutic method exists.* Now, it is immediately apparent that the expectation is that placebo controls should be limited to investigations where no proven treatment exists and, given that placebo controlled studies are often regarded as the drug development gold standard, this statement - if taken literally - is highly restrictive in most therapeutic areas. However from the perspective of paediatric drug development it opens up a separate area of controversy since, as described earlier, although plenty of drug treatments are used in children, most would hardly be described as proven. Indeed, if taken literally, the DH might actually be viewed as promoting the use of placebo controls to investigate many of the illnesses of childhood.

However Senn (2001) has challenged some of the basic principles stated in this 2000 revision to the DH and has proposed an alternate system of practical ethics based on the following principles:

I. The standard of care to which patients are entitled when not entered into clinical trials should be regarded as the standard by which the feasibility of the trial is judged.

II. Patients should not be entered into clinical trials if it involves them in an expected loss on any of the trial treatments compared with the standard they would get outside the trial unless the disease is not serious, the loss is temporary, and it has been explained to patients that such a loss is involved.

III. The trialist should always observe the fullest degree of consent practicable.

Senn's set of principles is actually aimed at encouraging a more practical approach to clinical trial design than that provided by the DH - particularly in relation to the use of placebo controls through his first two principles. In relation to current alternative treatments, Senn does not use the term "proven" but instead refers to the phrase *standard of care to which patients are entitled* in his first principle. This wording is quite interesting since it could actually prove to be more restrictive than the DH in relation to the use of placebo controls in paediatric trials. This would be the case if children were deemed to be entitled to off-label treatments and these treatments were considered to represent standard care. The Association of the British Pharmaceutical Industry (ABPI, 2001) states: *Many older medicines have not been tested on children, but experience over many years provides a sound base for their continuing and safe use*, which would appear to support the view that the current standard of care includes off-label treatments. Indeed as described in section 6.2, Banner (2002) considers it negligent not to regard off-label treatments as part of the paediatric pharmacopoeia. Support for off-label controls also comes from the FDA. As an alternative to placebo-controlled trials, the FDA has suggested that if an existing treatment were widely used but not proven in a particular indication, then this could be employed as a control but that the new treatment would be expected to demonstrate superiority (D Murphy of the FDA discussing solicited comments from the FDA's Anti-infective Drugs Advisory Committee/Pediatric Subcommittee of 23 April 2001 as reported in The Blue Sheet, 25 April 2001). This concept of requiring new treatments to demonstrate superiority to off-label controls will be re-visited later in this sub-section.

Now, since most treatments that have been tested in children over recent years have actually received regulatory approval for at least one new age category then perhaps

"proven in adults" is sufficient for a child to be entitled to treatment. Then again, the DH uses the phrase *current prophylactic, diagnostic, and therapeutic methods* and in practice the key issue with children in many disease indications will not be so much about the proven effectiveness of the drug but rather concerns regarding both the dose and formulation - that is, the method. In this respect while the treatment may have been demonstrated as being effective in adults perhaps it is the method that mostly remains unproven in children. As described earlier, under many circumstances approval of a drug for the treatment of children can be achieved through the simple extrapolation of adult clinical data to children via the conduct of an age specific PK study together with the accumulation of sufficient paediatric safety data. As such treatments that continue to be used off-label must either have not had the appropriate bridging studies conducted or the data must have been deemed inadequate to support a change to the label. Interestingly the FDA's Pediatric Advisory Committee (2000) has used yet another phrase in relation to the inclusion of placebo controls. The committee expressed the view that placebo controlled trials were acceptable if there were no *approved or adequately studied* therapies for children (Roberts, 2002).

The American Academy of Pediatrics (AAP, 1995) in their updated *Guidelines for the ethical conduct of studies to evaluate drugs in pediatric populations* give five conditions under which placebo controlled studies may be conducted. The first three of these conditions include *when there is no commonly accepted therapy for the condition* or the *commonly used therapy... is of questionable efficacy* or, where due to the safety profile, the risks associated may outweigh its benefits. The fourth condition relates to the acceptable use of placebo in add-on designs (an area that will be discussed later in this sub-section), while the fifth condition similarly permits placebo controls for chronic conditions whereby subjects have spontaneous exacerbations followed by periods of

remission. However the overriding feature is that the use of a placebo control group should *not place children at increased risk.*

Risk assessment is an important concept in the ethical conduct of clinical trials and in particular the weighting of the predicted harm and gain associated with participation. According to the American Academy of Pediatrics (AAP, 1995) benefits and risk must both be broadly defined when taking into account ethical considerations. For instance in relation to risks they suggest considering the following: *discomfort; inconvenience; pain; fright; separation from parents or familiar surroundings; effects on growth and development of organs; and size or volume of biologic samples.* However it is clear that whatever the risks and benefits identified, the weighting of these is very much an individual decision when it comes to trial participation. The role of the ethics committee therefore is simply to ensure that, for the study to proceed, a considered judgement is made on the basis of their considerable experience and expertise. To make this judgement, the ethics committee must themselves weigh the potential benefits of participation against the associated risks.

Senn (2001) uses the term *expected loss* to describe the result of the weighting calculation in terms of entering a subject into the trial regardless of the treatment assigned to the subject. The difficulty therefore with paediatric studies is that the expected loss will depend greatly upon one's view as to whether a drug has to be approved in the appropriate age category to be included in the calculation. If one takes the view that effective and safe treatments already exist even if they can only be used off-label then the expected loss may be non-trivial with a simple placebo control (unless in relation to an add-on design). However, if drug approval is a prerequisite then the expected loss of comparing a test drug to a placebo (or even to no treatment) may be zero or one might find an expected gain.

Included in Senn's second principle is the phrase *outside the trial* and therefore the loss function calculation must also include any benefit or detriment accruing from simple participation in a clinical trial when compared to the standard of care outside of the trial. For instance, the FDA is reported to believe that paediatric subjects actually receive a direct benefit from participating in placebo-controlled trials – the result of increased monitoring and enhanced care ((D Murphy of the FDA discussing solicited comments from the FDA's Anti-infective Drugs Advisory Committee/Pediatric Subcommittee of 23 April 2001 as reported in The Blue Sheet, 25 April 2001). Within a trial the balance can be shifted towards trial participation through the implementation of innovative methods. For instance, invasive procedures - such as the taking of blood samples - can sometimes prove restrictive to the approval of placebo controlled paediatric protocols since the risk can be seen to outweigh the benefits in the placebo group. However cholesterol, bilirubin and glucose levels can now all be measured non-invasively through the skin, rather than from blood samples (PR Newswire, 23 June 2001 & 13 June 2001). (Indeed placebo controlled studies have been reported in many varied paediatric indications including autism, asthma, attention deficit/hyperactivity disorder (ADHD), cystic fibrosis, epilepsy, ear infection, hypercholesterolemia, respiratory syncytial virus (RSV), reversible obstructive airway disease, psoriasis, to name just a few.) Regarding trial design, the FDA's Pediatric Advisory Committee (2000) provides some practical suggestions to limit the expected loss of participants, such as the use of DSMBs to permit early trial termination for serious or life threatening conditions. For less serious conditions it recommends the inclusion of individual subject discontinuation criteria - that is, early escape - to limit the exposure to ineffective treatments. The FDA has suggested the use of a randomised withdrawal designs to demonstrate long-term effectiveness of drug treatment when a long-term placebo-controlled trial would be unacceptable (D Murphy of the FDA discussing solicited comments from the FDA's Anti-infective Drugs Advisory

Committee/Pediatric Subcommittee of 23 April 2001 as reported in The Blue Sheet, 25 April 2001). Such a design has been employed in rheumatoid arthritis where a placebo controlled study was considered unethical. In this case, all subjects were treated with intravenous Enbrel for 3 months then randomised to placebo injections or Enbrel. If a relapse occurred (full blow flare up observed) whilst on placebo then the child was switched back to Enbrel (The New York Times, 11 February 2001)

For non-serious diseases, Senn's second principle allows subjects to enter a study so long as one expects any loss to be temporary and the subjects receive an explanation regarding the nature of this loss. This leads into a discussion of the concept of minimal or acceptable risk. Now, in the US, a mechanism - based on an original classification developed by the Department of Health and Human Services (DHHS) - has been established as part of the Children's Health Act of 2000 to approve clinical trials that traditionally would not have met ethics committee approval criteria. Trials can now be approved that:

1. Do not involve greater than minimal risk
2. Involve greater than minimal risk but offer direct benefit to individual subjects
3. Involve greater than minimal risk without the possibility of individual benefit but are likely to lead to greater knowledge about the subject's condition
4. Are not otherwise approved but represent an opportunity to understand, prevent, or alleviate a serious problem affecting child health or welfare.

In this respect, it is clear that paediatric studies can still be conducted if there is an expected loss - so long as it is small or is modest but can be traded off against the benefits. However within these assignments it is notable that there are terms that would benefit from greater clarity. For instance, it has already been identified (AAP, 1995) that ethics

committees require guidance on the definition of minimal risk. Furthermore for Category 3, the expanded text compounds the problem with the inclusion of the caveat that greater than minimal risk must only represent *a minor increase over minimal risk.* Continuing with Category 3 and the vague terminology, the information gleaned on the individual's condition must be of *vital importance* in relation to understanding or ameliorating the condition. Indeed, according to the DHHS's Secretary's Advisory Committee on Human Research Protection (SACHRP), ethics committees are frequently finding these four risk categories difficult to interpret and in some cases are selecting the lower categories when *not otherwise approvable* would be more appropriate (Washington Drug Letter, 28 July 2003)

Once the expected loss has been established for a paediatric study, and the ethics committee has approved the study, the next step is to communicate this information to potential trial participants and their parents or guardians. In practice it will be the parents or guardian of the child that receives and hopefully understands the explanation and it is these people who provide written consent while the child – depending upon age and understanding - is asked to provide assent. For the case of small children therefore the rationale decision-making process where the expected loss is evaluated on a personal basis is not possible. Also one concern is that parental consent does not necessarily equate to parental understanding. In the US, it has been reported that in paediatric oncology trials around one half of consenting parents did not realise that their children where randomly assigned to different treatments while a quarter did not understand that enrolling in a clinical trial would entail anything other than receiving standard cancer treatment (D Drotar speaking at the 2001 American Psychological Association meeting, reported in The Arizona Republic, 5 September 2001). It is perhaps because of such findings that the FDA's Pediatric Advisory Committee has expressed a preference for the inclusion of those

children in clinical trials who are able to give their assent to participate - although this should not preclude the development of age appropriate explanations. Senn alludes to the practical issues of consent his third principle.

It is now the time to return to add-on designs. In Senn's critique of the DH, he discusses these designs at length. In these studies subjects are randomised to a test treatment or placebo control, say, but both groups also receive a concurrent base or reference treatment. Senn uses these designs to illustrate the inadequacy of the DH wording in relation to use of placebo controls and to show that the use of placebo is indeed ethical even when proven therapeutic methods exist. However such designs raise some issues from a paediatric drug development perspective. Although paediatric add-on designs have been used in areas such as AIDS and Bipolar I Disorder, there are inherent risks if the base treatment (standard of care) is off-label. Drug interactions are less predictable in children compared with adults and according to Leeder (1996) certain drug-drug interactions may be quantitatively more important at one developmental stage compared to another. Furthermore, if the information on the base treatment is not deemed sufficiently adequate to introduce appropriate paediatric drug labelling, then from the drug-drug interaction perspective it could hardly be viewed as providing the basis for recommending the additional step of having a test treatment added on. As such, a cautious approach should be adopted when considering add-on designs at the early stage of the paediatric drug development programme although according to the FDA's Pediatric Advisory Committee (2000) it is generally acceptable to add-on placebo to *the standard of care*. They add a caveat that the study should normally include individual discontinuation criteria although this requirement suggests a concern in relation to lack of efficacy rather than the observation of unanticipated safety concerns.

Now, it is clear that, in relation to the choice of control group, the range of principles described in this sub-section above must be interpreted carefully when directed towards paediatric drug development. Moreover, many different phrases have been used in an attempt to describe the points of reference from which placebo should be judged if it were to be employed as a control group. These include: *proven prophylactic, diagnostic, and therapeutic method; standard of care to which patients are entitled when not entered into clinical trials; approved or adequately studied therapies; and commonly accepted/ used therapy*. However none of these really ties down whether the availability of off-label treatment precludes the use of placebo as a control. Although some of the terms would benefit from more precise wording to increase clarity, there is perhaps a stronger case for the introduction of alternate or additional principles that could be applied to account for the currently limited paediatric pharmacopoeia. For instance, perhaps it should be stated that the capacity to accumulate data on experimental treatments should not be restricted by the presence of off-label treatment? Furthermore, perhaps it should be stated that experimental treatments should not be penalised for failing to show statistical superiority in comparison with off-label treatments in cases where a placebo-controlled study is deemed unethical? Is it not unfair that the burden of superiority should reside with the experimental treatment, that has undergone rigorous testing for efficacy and toxicity, when the reference is unproven and potentially unsafe? In this instance approval could instead be granted according to the Schwartz, Flamant and Lellouch (1980) methodology discussed in Chapter Two. That is, the treatment comparison is reduced to a simple a decision making criterion whereby the treatment that is numerically superior is deemed successful (assuming a two-sided α set to 1 and β to 0). The study is then powered to control the γ error – defined as the probability of reaching a conclusion with the wrong sign - that is, the recommendation of an inferior treatment.

In practice, due to the inherent flexibility contained within the current guidance, ethics committees play a key role through their interpretation of the principles laid down, their evaluation of available information on alternatives to placebo and in their ultimate determination of whether a proposed design is ethical or not. In this respect whether it is appropriate to use off-label controls will depend upon the wider body of evidence available and it is difficult to be prescriptive, in this respect. Notwithstanding this, placebo controlled study designs are widely employed in practice in paediatric drug development and as a result the accumulation of information on paediatric drugs accelerates daily. Indeed as more information is obtained on the use of drugs in children through the use of adequate and well designed trials the issue of off-label prescribing will be reduced greatly and as a result some of the difficulties highlighted earlier in applying the general ethical framework will diminish accordingly.

6.5.2 Long term follow up

The follow up of subjects in paediatric clinical trials is an interesting topic owing to the potential for some treatments to affect child development in an adverse way. In particular, the detection of latent effects on growth and sexual maturation, the uncovering of events such as autism for which onset is difficult to establish and the ability to identify treatments which may be harmful genetic triggers. Historically a one-year follow-up period has been considered to be long-term but this is increasingly becoming to be regarded as inadequate and it has been reported that the FDA now plans to require a much longer follow-up of paediatric subjects. In some cases the FDA have requested 10 year follow up data in the form of five year data plus evidence that an infrastructure has been developed for longer term follow-up (D Murphy speaking at the FDA's Anti-infective Drugs Advisory Committee/Pediatric Subcommittee of 23 April 2001 as reported in The Pink Sheet, 7 May 2001). (This is indeed an area where the FDA can seem unreasonably strict. For instance

the FDA had requested a minimum six-month study in 35 girls with familial hypercholesterolemia but denied paediatric exclusivity as only five girls were strictly treated for six months; all received the treatment - 32 for 161 days or more. However in the FDA'a view the pharmaceutical company should have allowed for drop-outs and should have treated for at least 180 days (Health News Daily, 21 June 2001).) For drugs recently approved by the FDA under the Best Pharmaceuticals for Children Act's paediatric exclusivity programme, the Office of Pediatric Therapeutics is now required to review all adverse events reported for one year following the date that exclusivity was granted (FDA, 2003).

The potential scope of long-term follow-up is directed towards investigating the effects on maturation including growth velocities, academic performance, and the development of malignancies - although in general serious events should be identifiable within the first two years of therapy. For instance, there is evidence to support the view that the treatment of some childhood cancers - like leukaemia – increases the risk of future secondary malignancies, such as those pertaining to the breast, thyroid and brain (Neglia JP *et al*, 2001). Furthermore, it has been reported by the Institute of Medicine (2003) that around 25% of children who survive chemotherapy experience delayed severe or life-threatening adverse events impacting areas such as growth, fertility, heart function, muscle movement of cognitive activity (AP Online, 26 August 2003). Also the long term use of antibiotics for viral infections - particularly of the middle ear which have a high prevalence in young children - leads to bacterial resistance. However apart from the practical difficulties in observing subjects over a long period of time, a further problem is the confounding effects of environmental and lifestyle factors together with treatment and diagnostic advancements that have the potential to impact both underlying disease incidence and child development. For example, obesity has been on the increase in the US and

susceptible individuals are now developing type II diabetes earlier. Obesity itself can increase the risk of joint problems, asthma, hypertension and gall bladder disease. Traditionally children have not been screened routinely for diabetes, as the disease was uncommon in adolescents (F Diamond speaking at the 2nd Annual Conference on Obesity as reported in Associated Press Newswires, 17 May 2001). Chronic treatment usually provides the greatest concern from a safety perspective - although the wrong treatment at the wrong time can also have long term consequence. For instance in the chronic treatment of eczema there are long-term treatment concerns with oral and topical steroids (thinning of skin, growth retardation) and topical corticosteroids (cataracts, glaucoma) to consider (R McAlister speaking at the 58th Annual Meeting of the American Academy of Dermatology as reported in PR Newswire, 10 March 2000). In oncology, as patients with cancer survive longer then delayed effects of treatment become more important and potentially new events may be observed - some of these may occur in follow-up periods but some may also occur in adulthood. Identification of delayed effects is important so that the impact can be minimised or prevented. Suggestions in the report by the Institute of Medicine (2003) - an arm of the US's National Academy of Sciences - include developing a guideline for follow-up, linking specialist sites and primary physicians, raising awareness of late effects that threaten cancer survivors and increasing research to prevent late effects (AP Online, 26 August 2003).

Simply investigating mean changes in height and weight has on at least one occasion been deemed as being inadequate by the FDA who were concerned that the analysis of one and two year data had not accounted for differential growth expectations with regard to age and gender. In response the FDA reviewer suggested an approach whereby the height and weight of each child is measured through time with the resultant data compared with a standard growth chart. Predefined criteria would then be used to determine significantly

altered growth velocity on an individual subject basis (Pharmaceutical Approvals Monthly, 1 May 2001). Alternatively the resulting derived percentile data could be used to compare treatment groups using analysis of covariance. However, the use of growth charts (or school chart for performance) is not without problems since these are usually constructed from cross-sectional rather than longitudinal data and at best are applicable for short-term follow-up. Furthermore, there are no truly internationally accepted charts and most are country-specific – a particular problem for International clinical trials. Even within a country there is sometimes no single accepted standard. For instance, the Royal College of Paediatrics and Child Health formed an expert consensus group to review the situation in the UK. They reported that such charts such should be regarded as references and not standards that define an optimum growth pattern. They state that clinically, head circumference and weight is used most intensively in infants while height is used between the ages of 5 and 15 years. The group focussed on four references (Tanner and Whitehouse, Gairdner-Pearson, Buckler-Tanner and UK 1990) highlighting the different within country options available and considered which ones were most appropriate for different variables and age groups. There final recommendation was that for most clinical purposes the UK90 (Freeman *et al*, 1995) was the reference of choice (Royal College of Paediatrics and Child Health, 2002). In the US, Paediatric growth charts have been used since 1977 and have been developed (and revised in 2000) by the National Center for Health Statistics (2003). These are based on data originating from the National Health and Nutrition Examination Survey and have also been adopted internationally by the World Health Organisation (WHO). Given the limitation of growth references, simple analyses of the raw data – adjusted for baseline values, age and gender – may be the most appropriate approach in the randomised setting. However one has to be careful with missing data and imputation methods such as last observation carried forward have clear limitations. For instance an alternative approach could be to substitute data forward using

country specific growth references such that the subject remains on the same percentile of the curve following drop out. However this approach may also have limitations in areas such as asthma where it has been shown that although inhaled steroids negatively impacts height in the short-term, in the long-term the children catch-up with if conventional doses of the treatments are used (Doull, 2004). Body mass index (weight/height^2) requires careful consideration since the growth charts are not monotonic with age in children, and in this respect analysis of the derived percentile data represents the best approach.

Public sensitivity to the long-term (or delayed) effects of drug treatment in children should not be underestimated as illustrated by the current debate regarding the safety of the measles-mumps-rubella (MMR) vaccine. Introduced to the UK in 1988, MMR was implicated in 1998 by Dr Andrew Wakefield as a cause of autism (and inflammatory bowel disease) leading to a dramatic reduction in vaccine uptake. This was despite the results of a Finnish long-term follow-up study in 1.8 million subjects (who had received three million MMR doses) that showed no link and found that serious causally related adverse events were rare. The authors of the study concluded that the risks were greatly outweighed by the benefit of disease avoidance (Patja *et al*, 2000). Indeed autism is a useful vehicle to highlight many of the pitfalls and problems associated with assessing long term safety. Autism is more common in boys and generally appears before the age of three years. Affected children have trouble communicating and interacting with others - for instance they may not respond to their names, fail to make eye contact and engage in repetitive behaviour such as rocking and head-banging. In severe cases, children become aggressive or injure themselves. Autism has been recognised as a syndrome since 1943 but changes in diagnosis criteria – together with greater awareness – may have led to milder cases being identified over the years. These factors may be partly responsible for the increased disease prevalence that has been noted since the 1960s when the prevalence was

reported to be between 4 and 5 in 10,000 persons. In contrast, recent studies estimate the prevalence as 10 in 10,000 persons (Committee on Children with Disabilities, 2001). It has been suggested that the increase over the past decade in the UK is due to a change in the diagnosis of behavioural disorders. That is, while the number of children diagnosed with autism has increased per annum there has been a corresponding reduction in the number diagnosed with behavioural disorders. For instance, the inability to recognise faces may prove to be an objective early indicator of autism in children and may lower the limit of detection from 2 years of age to one (G Dawson speaking at the Annual Meeting of the Society for Research in Child Development (2001) as reported in Associated Press Newswires, 17 April 2001)

Now, it is clear that autism has a genetic component since the rate of autism is about 0.2% or less but for siblings of person with autism it increases to about 3% while for identical twins the rate is 60% or more (Associated Press Newswires, 28 January 2001). However, rather than a single flawed gene (as is the case with Huntington's disease or cystic fibrosis), it seems more likely that a combination of genes together with one or more environmental factors increases susceptibility. It has also been suggested that autism may be caused by a defect in metal metabolism that leads to impairment in brain development with resulting hypersensitivity to toxic environmental substances (WJ Walsh, A Usman and J Tarpey speaking at the American Psychiatric Association Annual Meeting 2001 as reported in PR Newswire, 10 May 2001)

Interestingly some vaccines (although not MMR) use thiomersal as a preservative which contains ethyl mercury, and since methyl mercury is known to cause traits similar to autism there was initial concern that this may be an environmental trigger. It also became apparent that the schedule of infant vaccination could lead to a cumulative exposure to

ethyl mercury that was in excess of the accepted threshold for methyl mercury. As a result, a review of the evidence was undertaken by the Global Advisory Committee on Vaccine Safety (GACVS) which advises the WHO. The GACVS concluded that since the half life of ethyl mercury was very much shorter (1.5 hours) compared with methyl mercury (1.5 months) then relative exposure would be much reduced, and as a result the current data did not support concerns that the ethyl mercury contained in vaccines was unsafe. The GACVS did however encourage further research in this area (WHO, 2003). Interestingly thiomersal is to be excluded from all new vaccines in the US.

In relation to vaccination, the Institute of Medicine Immunisation Safety Review Committee has suggested five factors that should be considered when investigating the link with autism. These are: evidence of causality; biological plausibility of the adverse event hypothesis; the likelihood of competing alternative hypotheses; the trade off between societal benefit and individual risk of vaccination; and the level of public concern about vaccines (K Stratton speaking at the Healthcare Resources and Service Administration's Advisory Commission on Childhood Vaccines (2000) as reported in Health News Daily, 8 December 2000). The debate is likely to continue for many years to come but it would come as no surprise to find that toxins of a metallic nature - from perhaps a number of sources – are among the environmental triggers.

Randomisation is a key component in vaccine trials. As early as 1954, the Salk poliomyelitis vaccine was evaluated in the US using a hybrid trial design (Francis *et al*, 1955) that incorporated what was at the time a controversial randomised component ((Meldrum, 1998). The trial was essentially split into two; a randomised design in over 600,000 children in the first three grades of school (that is, aged 6 to 9 years) who were randomised to either vaccine or placebo; and an observed control design in over one

million subjects where children in the second grade (aged 7 to 8 years) received vaccine while children in grades two and three were followed up as untreated controls. Sites from 11 US states participated in the randomised study while sites from 33 states participated in the observed control study.

The CHMP *Notes for guidance on the clinical evaluation of vaccines* (CHMP/VWP/164653/2005, 2005) reinforces the importance of the RCT in evaluating vaccines. In this respect the control group could represent placebo or another vaccine. As usual the greatest degree of blinding should be incorporated that is practicably possible. However the regulatory authorities do not necessarily require the demonstration of protective efficacy - for instance, in conditions such as diphtheria and tetanus where immunological data are known to be predictive of infection protection. In other cases protective efficacy is simply not practicable – for instance, smallpox which was declared eradicated by the WHO in 1980, diseases for which the incidence is too low (such as brucellosis and Q fever) or disease outbreaks that are unpredictable or short-lived. (Note that the CPMP has issued specific guidance for the development of second generation vaccines for smallpox: *Note for guidance on the development of vaccinia virus based vaccines against smallpox* (CPMP/1100/02, 2002). In this case the formation of a pock of appropriate size at the site of inoculation - that subsequently crusted over and scarred - was historically the correlate with protection.) Some topical areas are particularly challenging in terms of demonstrating protective efficacy – for instance, protection against anthrax where there is no established immunological correlate with protection but where there exists the spectre of 'bio-terrorism induced' disease on a mass scale. Farrington and Miller (2001) review the methods to evaluate vaccines in humans – including methods for post-registration studies.

Post-authorisation evaluation of vaccines is an important consideration for vaccines in terms of both efficacy and safety. Indeed the effectiveness of a vaccine can result from both direct protection of the individual and from indirect protection of the unvaccinated through herd immunity. (The effectiveness of a vaccine in the general population is governed by the uptake rate and if this is suitably high then the infectious agent will be unable to spread and then even those not vaccinated will be protected (Senn, 2003).) Consideration of efficacy also relates to duration of immunity. In terms of safety, one concern is simply that since such a large tranche of the community will be vaccinated, rare adverse events can translate into a not insubstantial frequency of individual occurrences. Furthermore since uptake is typically high and widespread within the target population, identifying controls retrospectively when potential safety concerns are raised is difficult. Pre-authorisation, the minimum requirement is that the CTD *should be sufficient to reliably determine the nature and frequency of local and systemic adverse events occurring at a frequency >1/1,000* (CHMP/VWP/164653/2005, 2005). However it is acknowledged that a CTD based only on immunogenicity studies is unlikely to be able to identify rare events.

Consequently there is a need to manage the introduction of new vaccines in a controlled manner to provide some framework for ongoing and future safety evaluation – that is, beyond the assessment of short-term events such as fever and injection site reactions. (Indeed the regulatory authorities have identified the need to have a *Pharmacovigilance Plan* to evaluate post-authorisation safety in a prospective manner (CHMP/VWP/164653/2005, 2005).) One potential design option is to use a "stepped-wedge" design. This design has been attributed Louis Molineaux, who introduced the concept around 25 years ago in the context of infant mortality and the assessment of anti-malarial treatments (Smith and Hayes, 1991). Essentially the idea is to randomise units to

vaccine in a stepwise manner such that each randomisation group has an incremental delay in receiving the vaccine. In effect, a crossover design where the time of crossover is randomised and the crossover is in one direction only - that is, no vaccine to vaccine. Typically the unit is a cohort of subjects – for instance, based on geographic region, investigator site, school etc., and in this respect there are similarities to cluster randomisation designs. In fact this design was used when the hepatitis B vaccine was introduced in the Gambia to 60,000 infants over a four-year period in the 1980s (Gambia Hepatitis Study Group, 1987). In this instance, there was a need to demonstrate the effect of vaccine on chronic liver disease, yet since the disease typically only manifests itself 20-30 years following infection, the quandary was how to generate controlled data without delaying mass vaccination for decades. A national surveillance system was implemented with the aim of identifying new cases of hepatocellular cancer - and other chronic liver diseases - over a follow-up period of 40 years, while the ethical dilemma of restricted vaccination was negated since the vaccine was expensive and insufficient vaccine would have been available in the short-term to treat all newborns. In terms of the randomisation procedure, every 3 months one of the existing 17 vaccination teams was selected at random (without replacement) to introduce the vaccine to their specific region, and the procedure continued until all teams had been selected. (The randomization was also stratified by ecological zone). In terms of analysis it was envisaged that for the first 3 month period, the 1^{st} cohort would be compared with the 2^{nd} – 17^{th} cohorts. Similarly for the 3-6 month period, the 1^{st} – 2^{nd} cohorts would be compared with the 3^{rd} – 17^{th} cohorts, etc. (Other potential analysis methods could include mixed models and generalized estimating equations.)

The staggered introduction of vaccines in developed societies would be more controversial due to issues surrounding the denial of vaccine protection. However, vaccine uptake has

decreased in recent years – in particular following the controversy surrounding MMR. In a society where increasing numbers of parents are becoming more informed about clinical research and making their own judgements in terms of risk/benefit, it could be that a program of staggered vaccination would be deemed morally acceptable by many – particularly for very rare diseases, or those only likely to appear through acts of bioterrorism. It might also be appropriate for vaccines against generally non life-threatening diseases such as chickenpox. In this respect although chickenpox vaccines exist and are routinely used in the US, their introduction in the UK has been much more more limited. Although such vaccines might prove popular if a mass programme were introduced in the UK, uncertainty over the risk/benefit ratio might provide an environment where staggered introduction was ethically acceptable using GP surgeries or regional health authorities.

A non-randomised approach sometimes used to investigation of safety of vaccines is the use of case control designs. In this respect these are effectively a comparison of vaccine adopters versus non-adopters (Kirkwood *et al*, 1997). However confounding is an issue as the groups may differ in many key aspects. Such investigations are also more suited to the investigation of specific events rather than multi-factorial safety investigations and are more retrospective in nature.

In summary, the primary tool available for assigning cause to effect is randomisation - even for delayed adverse events. With so many other factors potentially impacting long term outcome, the use of uncontrolled investigation is seriously flawed. Controlled investigation provides baseline data and prospective monitoring using precisely defined tools that enable direct comparisons to be made. Of course due to the almost infinite type of unrelated events that could occur once treatment has commenced - and in acute cases

once treatment has completed - false positive findings are a distinct possibility and the subsequent responsibility to investigate the biological plausibility of such findings should not be underplayed. However alternate solutions can only ever be presented as poor imitations of a casual solution.

6.6 DISCUSSION

Although paediatric drug development provides a series of unique challenges, the underlying requirements for well designed, well controlled and well conducted clinical investigation, remain key. Randomisation remains the best tool at the researchers' disposal to assign cause to effect and, in the absence of alternate approved treatments, placebo has a key role to play in determining the absolute effect of the test treatment and in advancing knowledge in relation to the paediatric pharmacopoeia. In many ways it is simply the case that the challenges of conducting drug development in adults are accentuated - doses levels and formulations need more careful thought. Furthermore the choice of control is more complicated while study conduct and follow-up is prolonged. Wilson (1996) lists four principles that all paediatric strategies should follow and it is difficult to argue with his view. Wilson's principles are: if a drug is to be used in children, then it must be tested in them; a dose for the child is central to paediatric therapeutics; immature clearance impacts on drug dose and hence on efficacy and toxicity in children; and clinical investigation of drugs in pediatrics applies a controlled trial in the study design. However from a pragmatic perspective, a key success factor in paediatric drug development is obtaining access to both paediatric experience and knowledge and also to the associated infrastructure. Driven from the US, a paediatric infrastructure is now spreading and the hope is that the therapeutic orphan will soon find a home. Indeed as a sign of the changing times, even thalidomide is being investigated again and this time in the paediatric

population - the indication, a rare immune system disease called Langerhans Cell Histiocytosis (Associated Press Newswires, 28 February 2001).

CHAPTER SEVEN: DISCUSSION

Lily the Pink, she
Turned to drink, she
Filled up with paraffin inside
And despite her medicinal compound
Sadly Picca-Lily died.

Up to Heaven
Her Soul ascended
All the church bells they did ring
She took with her medicinal compound
Hark the herald angels sing.

7.1 INTRODUCTION

In the following section of this Discussion chapter, a problematic therapeutic area has been selected to illustrate some of the practical challenges faced when addressing sub-populations and subgroups. This chosen area is the treatment of subjects with febrile episodes of neutropenia. In this section, the findings and ideas developed during the course of this Research Thesis will be used to offer solutions that are regulatory compliant and that would enable reliable and robust conclusions to be drawn from the clinical trial data. These solutions will cover both design and analysis. In the third section of this chapter, modifications to the text of current regulatory guidance will be proposed based on the observations made in earlier chapters – in particular chapters 4 and 5. In this respect the aim is to identify omissions, improve consistency and increase clarity. The fourth section looks to the future and the challenges facing drug developers in the areas of sub-populations and subgroups – specifically in relation to the use of genetic information. In the fifth and final section, some concluding thoughts will be offered in relation to the generalisation of clinical trial data to clinical practice. In this respect individualised treatment will be contrasted to universal remedies.

7.2 A CASE STUDY: NEUTROPENIA

Febrile neutropenia is a particularly difficult therapeutic area and, from a statistical perspective, is fraught with many of the challenges highlighted earlier in this Research Thesis in relation to analysis populations, subgroups and the control of potential bias. In this section the aim is to identify some of the challenges and to propose tactics for valid and robust statistical analyses.

7.2.1 Background

Neutropenia is characterised by haematological abnormalities in the blood due to underlying disease, treatment regimen or congenital abnormality. Subjects are formally diagnosed as being neutropenic if they have <500 polymorphonuclear leukocytes/mm^3 in their blood or if the neutrophil count is between 500 and 1000 neutrophils/mm^3 but the count is expected to drop to <500 because of antecedent therapy (Hughes *et al*, 1992). Such subjects are susceptible to infection and a febrile episode in the neutropenic subject is defined as the presence of fever - that is, temperature >38.3 °C. Broad-spectrum anti-infective treatments (often given via the intravenous route) have been found to reduce morbidity and mortality markedly - despite the fact that around 50% to 75% of cases are categorised as *fever of unknown origin* where a pathogenic cause is never identified (Hughes *et al*, 1992). That is, no pathogen is detected in the various blood samples taken from the subject. With regard to confirmatory clinical trials, it is standard practice to employ an active control group due to the serious nature of the disease and these studies are designed to show non-inferiority of the test treatment to a reference treatment. Both clinical and bacteriological responses are considered important and both are reduced to a straightforward dichotomous outcome – that is, treatment success or failure – for the purpose of treatment comparison.

With regard to guidance, the Report of a Consensus Panel from Immunocompromised Host Society (Consensus Panel IHS, 1990) makes particular interesting reading from a statistical perspective. This report published in 1990 was an attempt to address standards in *the design, analysis and reporting of clinical trials on the empirical antibiotic management of the neutropenic patient.* The report was subsequently used as the basis for the document entitled *General guidelines for the evaluation of new anti-infective drugs for the treatment of febrile episodes in neutropenic patients* (Hughes *et al*, 1992) which was sponsored by the FDA. Although broader anti-infective guidance was later published, these documents provide an interesting insight into the contrast between clinical and statistical thinking – much of which still remains today.

7.2.2 Blinding

Whilst recognising the value of blinding the IHS did not regard blinding as a mandatory requirement – taking the view that this was often complicated and impractical, and if used required an unblinded observer to monitor safety. As highlighted above, studies in febrile neutropenia employ an active control group rather than placebo - although perhaps there has been under use of placebo add-on studies. Blinding is indeed a challenge in these studies since competing treatments are often given intravenously and dosing frequency can differ. However blinded or partially blinded studies are possible, and it is important that safety is assessed without knowledge of the treatment received, whilst ensuring that provision exists for emergency unblinding. If for practical reasons a blinded study cannot be undertaken, central randomisation is key to ensuring control of selection bias in the assignment of subjects to randomised treatment.

7.2.3 Factors known to influence outcome

The IHS identifies the degree and duration of neutropenia as affecting the risk of infection but actually recommends stratification by underlying cancer (leukaemia versus sold tumour) and age (paediatric versus adult subjects) since these two variables are associated with degree and duration and are more reliably recorded pre-randomisation. Other potential factors are also indicated, although baseline pathogen presence and species - although naturally appealing – are impractical since these are not usually known prior to the start of randomised treatment (as discussed in Chapter Two). The IHS warns against over-stratification and a particular problem is stratification by study centre since individual centres frequently randomise just a few subjects. Another factor to consider is that confirmatory studies for world-wide drug registration usually require centres from multiple countries and in these different countries the type and resistance potential of pathogens varies. (Indeed this is the very reason for conducting multi-centre studies in neutropenia - that is, to ensure that the conclusions are geographically applicable in a broad sense – and in recent years the advent of Good Clinical Practice (ICH E6, 1996) in developed countries has even led to the FDA agreeing to accept pivotal data generated outside the US. In fact it is now commonplace in the European Union to conduct pan-European studies extending to countries such as Israel, Russia, Poland, Turkey – and even as far as to Australia and South Africa. In some cases, US and European centres are included in the same study although it is rare to find countries from the Far East – such as Japan and China – combined with European or US centres in the same trial.) Hence one appropriate design tactic is to stratify a study by underlying cancer and country but not centre. (Note that Paediatric studies are usually conducted separately and require additional stratification by age category as discussed in Chapter Six.)

7.2.4 Randomised more than once

One of the more bizarre recommendations of the IHS is the encouragement of the inclusion of multiple episodes of febrile neutropenia originating from the same subject in a clinical trial. That is, it recommends protocols that allow subjects to re-enter the trial and be re-randomised an unlimited number of times. Their view is that each episode of neutropenia is clinically independent of the last. For instance, neutropenia is often associated with the effects of chemotherapy in cancer, where each cycle of chemotherapy has a short-term impact on the immune system resulting in increased vulnerability to bacteria. In this respect, an infection may occur before a subject's immune system recovers but the invading species of bacteria may well differ from episode to episode. Although clearly independent to some extent, factors such as the pharmacokinetic profile of the drug in the subject, susceptibility to toxic events, and the simple fact that the subject survived the first episode, is evidence enough that the response second time around will not be independent of the first result. (Indeed in one study the author was involved with, a subject was re-entered 8 times giving 9 separate randomisations!). However from a clinical perspective, a study that does not allow the re-entry of subjects generates independent data but fails to represent true clinical practice. To reconcile the clinical and statistical perspectives some solutions have been proposed including the re-entry of subjects without re-randomisation. In this case the subject receives the same randomised treatment each time but only the first randomised episode is included and the primary analysis population – subsequent treated episodes providing valuable information on consistency of response and the development of bacteriological resistance. Other options include randomising subjects to a sequence of treatments such that re-entered subjects cross-over providing in addition an informal subset of subjects who have received both treatments.

7.2.5 Primary analysis populations/sets

In addition to the issues discussed in Sub-section 7.2.4, neutropenia - and indeed anti-infective indications in general - raise some interesting challenges regarding the definitions of analysis populations (or to use the ICH E9 term, analysis sets). For Clinical response, the primary analysis population is reasonably straightforward and according to the arguments presented in Chapter Five for non-inferiority, the ITT principle can be used (as described in Chapter Two) to define the Full Set. In this respect, the analysis population would be the strict "all subjects randomised, as randomised".

In Chapter Five, a recommendation was made regarding the use of an eligible analysis population for non-inferiority and this approach is directly applicable to the corresponding Bacteriological response which is derived from the individual pathogen data. Although blood samples are taken from subjects prior to randomisation in febrile neutropenia, the results are not usually available for 48 hours (for it takes this long to culture the bacteria). It is not uncommon therefore to find that no pathogen has been cultured for a randomised subject and as such it is impossible to assign a bacteriological response to treatment. (This can be further complicated if subjects receive prophylactic antibiotics prior to randomisation since these can mask a pathogen present in the sample and can increase the risk that no baseline pathogen is identified.) This very issue was the subject of world-wide debate in the early 1990's during which time there was growing recognition of the need to consider clinical and bacteriological responses to treatment separately. As a result a consensus was reached that clinical outcome should be evaluated using all randomised subjects using the ITT principle, while bacteriological response would be evaluated in the sub-population of subjects who had at least one pathogen identified from the baseline sample (sometimes referred to as a modified ITT population). The use of the ITT principle for clinical outcome recognised the fact that if in practice subjects were treated on the basis

of clinical signs and symptoms of infection then there was a need to answer the pragmatic question as to how the treatments compared in practice. In contrast the analysis of bacteriological outcome was viewed as an explanatory question - comparing treatments with regard to microbiological effect. Interestingly, Gillings and Koch (1991) have recommended a further restriction in relation to resistant pathogens. According to their position, including subjects in an ITT analysis with pathogens known to be resistant to randomised treatment is not useful. However it is not clear why Gillings and Koch hold this view since resistance is clearly treatment related and subjects are usually randomised prior to having knowledge of the type of pathogen present. Furthermore, one could argue that restricting the analysis population to only those subjects who have pathogens present that are susceptible to both study treatments (test and reference) is to some extent uninformative since both treatments should perform well in these cases. Indeed it is the very fact that some newer treatments have a broader spectrum of activity than older treatments that is of interest in the early treatment of high-risk subjects in whom a pathogen has yet to be identified.

7.2.6 Protocol violations and missing data

Due to the complicated nature of these clinical trials, procedural non-compliance, subject withdrawal and missing data are relatively common. In particular, subjects who are not responding to randomised treatment will tend to have another antibiotic added by the Investigator within a short timeframe. In this eventuality, subjects are regarded as treatment failures according to protocol. Another feature is that although short-term response (72 hours post randomisation, say) is evaluated, long-term response is considered to be of greater importance, with the increased risk of loss to follow up. Inevitably therefore, a procedure needs to be employed to account for, and handle, missing data. One approach has been to categorise subjects with missing outcomes as failures and, from

Chapter Two (Sub-section 2.3.3), it can be seen that this approach will lead to an increase in bias in the presence of differential misclassification when $\pi_t = \pi_r = \pi$ and $\pi \to 1$ (as is the case in anti-infective studies), since $E(p_t' - p_r') = \pi(\theta_r - \theta_t)$. However, from a regulatory perspective, this may be viewed as advantageous since it is likely to maximise the difference between treatments and may make it harder to demonstrate non-inferiority – unless of course the misclassification rate is higher in the reference treatment group, in which case the bias will actually favour the test treatment. An alternative approach is to carry forward the response assigned at short-term follow-up to long-term follow-up. This again highlights the importance of blinding to ensure that the long-term follow-up of subjects is unbiased, and the need for sensitivity analyses to ensure that the conclusions are largely unaffected by the data conventions chosen.

7.2.7 Statistical analysis - including interactions and subgroups

In sub-section 7.2.3, underlying cancer (leukaemia versus sold tumour) and age (paediatric versus adult subjects) were identified as baseline factors that are known to influence the risk of infection - although the expectation is that, in practice, paediatric subjects would be the subject of a separate study. Country was also highlighted as a potential source of influence as bacterial resistance patterns can vary from country to country. Furthermore, although it was shown that *a priori* it was not possible to stratify by pathogen status (at least one identified pathogen versus no pathogen identified), retrospective stratification is feasible. Indeed it would be expected that the response to treatment would be influenced by whether a pathogen was confirmed to be present or not. The objective of the study would most likely to be to demonstrate non-inferiority of the test treatment to the reference treatment and a non-inferiority margin would have been pre-selected to compare the treatments with respect to the proportion of successes or cures. As shown in Chapter Five, the most appropriate model to compare treatments is the odds ratio formulation, since

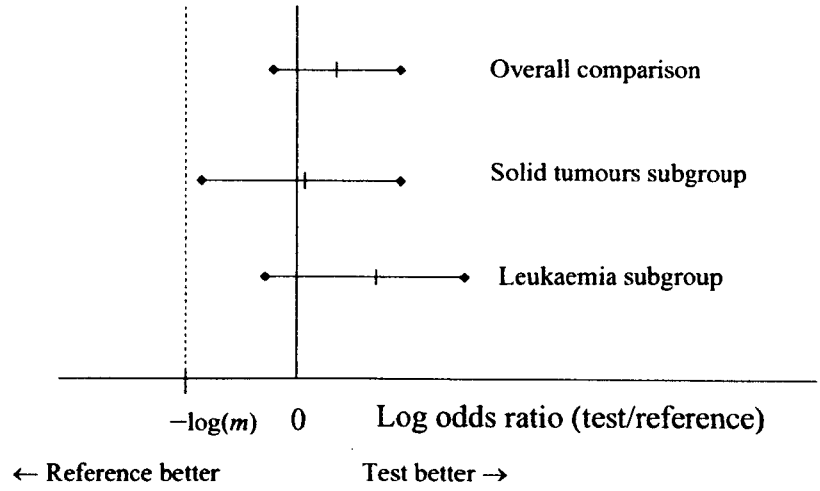
conditioning on covariates is straightforward and the log odds ratio is not bounded like the difference in proportions. Therefore using a margin specified in terms of the log odds ratio the primary analysis would be performed using a logistic model including the design features country and underlying cancer - and potentially pathogen status also for Clinical response. Non-inferiority would then be investigated by comparing the lower 95% confidence limit for the odds ratio with the selected margin and if the limit were greater than the margin then non-inferiority would be claimed. (Refer to Chapter Five for details of the underlying statistical methodology.)

The primary investigation of the consistency of treatment effect within pre-specified subgroups – that is, underlying cancer and country – would be achieved by fitting a treatment by factor interaction term to the model - separately for each factor. Now, although the overall study objective is non-inferiority, the test for the difference between the treatment differences essentially has a null hypothesis of no interaction. However as documented in Chapter Four, a preferable approach when investigating treatment by subgroup interactions might be to apply a symmetric equivalence margin for the interaction parameter. For instance a margin $\frac{\log(m)}{2}$ equal to one half of the selected non-inferiority margin, m, on the log scale. Now, the interaction between treatment and underlying cancer would be quite straightforward since the factor has just two levels. However for the interaction between treatment and country, the number of levels could be quite large and an arbitrary base country would be required to enable estimates of the parameters to be constructed. As suggested in Chapter Four (Figure 4.1), it would be informative to present the data graphically (estimate of interaction parameter with 95% confidence interval versus the two interaction margins). For each estimate, if the confidence interval did not include zero then the given contrast would, of course, be statistically significant at the two-sided 5% level. (Note, that the likelihood ratio or standardised range test, could be used to

determine whether a statistically significant interaction were qualitative and quantitative.) However, as suggested in Chapter Four, a more informative approach might be to calculate the posterior probability that the interaction parameter is contained within the symmetric margins$\left(-\frac{\log(m)}{2},+\frac{\log(m)}{2}\right)$. This would show the level of support for the conclusion that the treatment effect is consistent between pairwise levels of the factor.

To supplement these interaction analyses, it would also be appropriate to construct a 95% confidence interval for the log odds ratio within each subgroup and to present this alongside the overall treatment comparison with the non-inferiority margin as a point of reference. (Note that the estimate and confidence interval for the overall treatment comparison should be adjusted for the factor concerned since, as shown in Chapter Five, the unadjusted estimate of the log odds ratio will tend to be diluted towards zero if the factor impacts outcome.) Figure 7.1 shows how this might be presented. Now, if the lower confidence limit for each subgroup is actually >-log(m), then this would indeed provide robust evidence that the test treatment was non-inferior to the reference treatment in all subgroups. However this is an unrealistic expectation – particularly as the number of levels of the factor increase. The directional advantage approach discussed in Chapter Two is an alternative tactic or perhaps even a second level step in determining the degree of robustness. In this case, each subgroup must satisfy the directional advantage criterion such that the point estimate must be >-log(m); as is the case in Figure 7.1.

Figure 7.1. Presentation of subgroup data for non-inferiority study (log odds ratio and 95% confidence interval).



In the next section of this chapter, the observations of earlier chapters – in particular chapters 4 and 5 - will be considered in relation to modification of specific sections of current regulatory guidance.

7.3 GUIDANCE AMENDED

Throughout this research thesis, reference has been made to regulatory guidance and in a number of instances, omissions and inconsistencies have been highlighted. In this section modifications to the text of specific sections of the statistics regulatory guidelines will be proposed (bolded text), beginning with ICH E9: *Statistical principles for clinical trials*. In this case proposed modifications relate to the impact of covariate adjustment on estimate precision for the broader class of generalised linear models together with an acknowledgement of the impact on non-inferiority trials, the role of analysis sets, and the importance of confidence intervals in the interpretation of interactions.

Section 1.2: Scope and Direction

CURRENT: Many of the principles delineated in this guidance deal with minimising bias (see Glossary) and maximising precision.

PROPOSED: Many of the principles delineated in this guidance deal with minimising bias (see Glossary) and maximising **efficiency**.

Section 5.2: Analysis Sets

CURRENT: Decisions concerning the analysis set should be guided by the following principles: 1) to minimise bias, and 2) to avoid inflation of the type I error.

PROPOSED: Decisions concerning the analysis set should be guided by the following principles: 1) to **control the direction of** bias, and 2) to avoid inflation of the type I error.

Section 5.2.3: Roles of the Different Analysis Sets

CURRENT: The full analysis set and the per protocol set play different roles in superiority trials (which seek to show the investigational product to be superior), and in equivalence or non-inferiority trials (which seek to show the investigational product to be comparable, see section 3.3.2). In superiority trials the full analysis set is used in the primary analysis (apart from exceptional circumstances) because it tends to avoid over-optimistic estimates of efficacy resulting from a per protocol analysis, since the non-compliers included in the full analysis set will generally diminish the estimated treatment effect. However, in an equivalence or non-inferiority trial the use of the full analysis set is generally not conservative and its role should be considered very carefully.

PROPOSED: In superiority trials the full analysis set is used in the primary analysis (apart from exceptional circumstances) because it tends to avoid over-optimistic estimates of efficacy since the non-compliers included in the full analysis set will generally diminish the estimated treatment effect. However, in an equivalence or non-inferiority trial (which seek to show the investigational product to be comparable, see section 3.3.2) the use of the full analysis set is generally not conservative and its role should be considered carefully. **In particular the exclusion of subjects from the full analysis set on the basis of pre-randomisation violations – such as**

subjects without the specified disease severity – should be considered to further the opportunity to detect true treatment differences. The per protocol set is typically used in a supportive role for both superiority and equivalence or non-inferiority trials since it is open to bias through subject exclusion – the direction of which is difficult to predict.

Section 5.3: Missing Values and outliers

CURRENT: None

PROPOSED: Addition of text from the *Points to Consider on Missing Data* (CPMP, 2001): **It is considered of particular importance to ensure that the selected method is a conservative approach and does not favour the study's working hypothesis (intentionally or unintentionally).**

Section 5.5: Estimation, Confidence Intervals and Hypothesis Testing

CURRENT: A description should be given of any intentions to use baseline data to improve precision or to adjust estimates for potential baseline differences, for example by means of analysis of covariance.

PROPOSED: A description should be given of any intentions to use baseline data to improve **relative efficiency** or to adjust estimates for potential baseline differences. **Careful consideration in the randomised setting should be given to the distinction between models (such as analysis of covariance) where covariate adjustment tends to increase the precision of the estimate and those (such as logistic regression) where precision is decreased, but efficiency is increased overall.**

Section 5.7: Subgroups, Interactions and Covariates

CURRENT: Pre-trial deliberations should identify those covariates and factors expected to have an important influence on the primary variable(s), and should consider how to account for these in the analysis in order to improve precision and to compensate for any lack of balance between treatment groups.

PROPOSED: Pre-trial deliberations should identify those covariates and factors expected to have an important influence on the primary variable(s), and should

consider how to account for these in the analysis in order to improve **efficiency** and to compensate for any lack of balance between treatment groups.

Section 5.7: Subgroups, Interactions and Covariates

CURRENT: The treatment effect itself may also vary with subgroup or covariate – for example, the effect may decrease with age or may be larger in a particular diagnostic category of subject. In some cases such interactions are anticipated or are of particular prior interest (e.g. geriatrics), and hence a subgroup analysis, or a statistical model including interactions, is part of the planned confirmatory analysis. In most cases, however, subgroup or interaction analyses are exploratory and should be clearly identified as such; they should explore the uniformity of any treatment effects found overall. In general, such analyses should proceed first through the addition of interaction terms to the statistical model in question, complemented by additional exploratory analysis within relevant subgroups of subjects, or within strata defined by the covariates.

PROPOSED: The treatment effect itself may also vary with subgroup or covariate – for example, the effect may decrease with age or may be larger in a particular diagnostic category of subject. In some cases such interactions are anticipated or are of particular prior interest (e.g. geriatrics), and hence a subgroup analysis, or a statistical model including interactions, is part of the planned confirmatory analysis. In most cases, however, subgroup or interaction analyses are exploratory and should be clearly identified as such; they should explore the uniformity of any treatment effects found overall. In general, such analyses should proceed first through the addition of interaction terms to the statistical model in question, complemented by additional exploratory analysis within relevant subgroups of subjects, or within strata defined by the covariates. **For both planned and exploratory analyses, confidence intervals are an important aid to the interpretation of subgroup and interaction analyses.**

Section 5.7: Subgroups, Interactions and Covariates

CURRENT: None

PROPOSED: **In contrast to analysis of covariance models, the inclusion of covariates associated with the outcome for some commonly used generalised linear models (e.g. logistic and proportional hazards models) will tend to decrease the precision of the estimated treatment difference rather than increase it. However efficiency will tend to increase since the corresponding estimate of the treatment difference will tend to increase and more than counterbalance the corresponding decrease in precision.**

Addition of the following text from the *Points to Consider on Adjustment for Baseline Covariates* (CPMP,2003). **In such models the adjusted parameters and unadjusted parameters have different interpretations: it is essential that in any presentation of adjusted analyses, the applicant clearly and precisely explains the meaning of the estimated effect size.**

Careful consideration should be given to the use of such models in non-inferiority designs where a decrease in precision but increase in the estimate may impact the interpretation of the confidence limits in comparison with an unadjusted analysis.

The second regulatory guideline considered is the *Points to Consider on Adjustment for Baseline Covariates* (CPMP, 2003) where proposed modifications relate to the absence of a reference to non-inferiority designs, the impact of covariate adjustment on estimate precision for the broader class of generalised linear models, and the use of confidence intervals to interpret interaction analyses.

Section II.1: Association with the Primary Outcome

CURRENT: Adjustment for such covariates generally improves the efficiency of the analysis and hence produces stronger and more precise evidence (smaller p-values and narrower confidence intervals) of an effect.

PROPOSED: Adjustment for such covariates generally improves the efficiency of the analysis and hence produces stronger evidence (smaller p-values) of an

effect.

Section III.1: General considerations

CURRENT: The nature and the number of covariates included in the analysis may affect the interpretation of the analyses, especially in non-linear models. In such models the adjusted parameters and unadjusted parameters have different interpretations: it is essential that in any presentation of adjusted analyses, the applicant clearly and precisely explains the meaning of the estimated effect size.

PROPOSED: The nature and the number of covariates included in the analysis may affect the interpretation of the analyses, especially **for the broader class of generalised linear models – including logistic regression and Cox-regression.** In such models the adjusted parameters and unadjusted parameters have different interpretations: it is essential that in any presentation of adjusted analyses, the applicant clearly and precisely explains the meaning of the estimated effect size.

For logistic and Cox models, adjustment for covariates associated with the outcome will tend to decrease the precision of the estimate of treatment difference although efficiency will tend to increase overall since the corresponding estimate of the treatment difference will increase. Careful consideration should be given to non-inferiority designs where a decrease in precision but increase in the estimate may impact the interpretation of the confidence interval in comparison with an unadjusted analysis.

Section IV.3: Treatment by covariate interaction

CURRENT: Tests for interaction often lack statistical power and the absence of statistical evidence of an interaction is not evidence that there is no clinically relevant interaction. Conversely, an interaction cannot be considered as relevant on the sole basis of a significant test of interaction. Assessment of interaction terms based on statistical significance tests is therefore of little value.

PROPOSED: Tests for interaction often lack statistical power and the absence of statistical evidence of an interaction is not evidence that there is no

clinically relevant interaction. Conversely, an interaction cannot be considered as relevant on the sole basis of a significant test of interaction. Assessment of interaction terms based on statistical significance tests is therefore of little value **and confidence intervals should be used to aid clinical interpretation.**

The third regulatory guideline considered is the *Points to Consider on Switching between Superiority and Non-inferiority* (CPMP, 2000) where proposed modifications relate to the choice of analysis sets.

Section IV.1.4 and Section IV.2.3: Choice of analysis set

CURRENT: In a superiority trial the full analysis set, based on the ITT (intention-to-treat) principle, is the analysis set of choice, with appropriate support provided by the PP (per protocol) analysis set. In a non-inferiority trial, the full analysis set and the PP analysis set have equal importance and their use should lead to similar conclusions for robust interpretation. A switch of objective would require this difference of emphasis to be recognised.

PROPOSED: **Since the 95% confidence interval for the treatment difference is central to the interpretation of the trial, and once the data are observed it is only the conclusion that that may change and not the confidence interval itself, it is important to address prospectively the statistical conventions (e.g. handling missing data) used to construct the confidence interval. Typically it is important to ensure that the selected conventions produce a conservative approach that does not favour the study's working hypothesis. However switching the objective of the comparison from non-inferiority to superiority (or vive versa) leads to a juxtaposition that must be recognised.**

In a superiority trial the full analysis set, based on the ITT (intention-to-treat) principle, is the analysis set of choice **while in a non-inferiority trial the use of the full analysis set is generally not conservative and its role should be considered carefully. In particular the exclusion of subjects from the full analysis set on the basis of pre-randomisation**

violations – such as subjects without the specified disease severity – should be considered to further the opportunity to detect true treatment differences. The per protocol set is typically used in a supportive role for both superiority and non-inferiority trials since it is open to bias through subject exclusion – the direction of which is difficult to predict.

Carefully chosen analysis sets should be selected anticipating a possible switch in objective, and confidence intervals produced from these sets should lead to similar conclusions for robust interpretation.

The *Points to Consider on Missing Data* (CPMP, 2001) and *Points to Consider on Multiplicity issues in Clinical Trials* (CPMP, 2002) are considered adequate and no text modification is proposed. The *Points to Consider on Application with 1.) Meta-analyses and 2.) One Pivotal study* (CPMP, 2001) is less relevant to this research thesis and is not considered in the context of the observations made in earlier chapters.

In the fourth section of this chapter the focus switches to future challenges facing drug developers - in particular the investigation of treatment differences based on genetic make-up.

7.4 THE BRAVE NEW WORLD OF GENETICS

The next logical step with regard to the subdivision of subjects and the investigation of the consistency of effect between subgroups is inevitably the use of genetic information in drug development. In this respect it is important to distinguish pharmacogenetics - which investigates the potential for interaction between specific genes and drug treatments - from pharmacogenomics - which targets the inheritable response to drugs over the entire genome. According to Reidenberg (2003), *Science is continuing to take patients with a disease, to stratify them into smaller and smaller groups of increasingly more*

homogeneous individuals, and then to develop drugs specific for these smaller groups of more homogeneous individuals. Such homogeneity implies genetic homogeneity - well at least on a specific gene or two! For instance, it is has been estimated that genetic differences that encode the metabolising enzymes, transporters and targets of drugs may produce between 20% to 95% of the variation in drug response between subjects – although typically it is the actual interplay between many genes that is key (Evans and McLeod, 2003). However according to Senn (1999), *we commonly underestimate pure random within patient variation and such within-patient variation cannot, by definition, be genetic* – and in order to identify the relevant components of variation, cross-over designs using multiple periods must be used. Indeed with parallel group designs Senn (2001) shows that it is impossible to distinguish the main effect of subject from both the subject by treatment interaction (in effect the upper bound for the genome by treatment interaction) and the within subject error. (Investigating subgroup effects (sex or a factor based on a specific gene, for instance) can separate out some of the variation in relation to subject and the subject by treatment interaction, however.)

The real goal of the many proponents of such genetic sub-division is in fact individualised treatment; a Promised Land where specific drugs are developed for specific patients and where all effects are known - efficacy without toxicity, war without tears. Perhaps, even a land without statisticians, where the likelihood, $P(evidence \mid hypothesis)$, becomes purely deterministic [$P(e \mid h) = 1$] and there is no need for parallel group designs let alone ones of the multi-period, cross-over variety. However the element that is frequently lost in this debate is best illustrated by considering the challenges faced when developing drug treatment for children. As described in Chapter Six, notwithstanding the formulation issues, the appropriateness of a particular drug regimen varies enormously according to the age and developmental maturity of the subject. Although this variability is at it greatest

when the child is very young, the modification of dose and indeed treatment continues into old age – when liver, renal and other functions deteriorate. Drugs also interact with others drugs and concurrent diseases – they interact sometimes with what you eat and when you eat it. Physical activity, exposure to infectious diseases and viruses can all have an impact. However during a life a constant physiological change and environmental bombardment the one thing that remains a constant is your genetic make-up. In essence therefore, drugs treatments are already somewhat tailored throughout life in the absence of genetic variability, and the use of genetic information simply represents a further opportunity for refinement. Indeed if any technique can be used to tailor treatment to individual subjects it is the use of pharmacokinetic monitoring in these treated subjects. A couple of other notes of caution are required for those seeking the Promised Land. Firstly, since cause and effect are related to how quickly a drug is absorbed into the body and how quickly - and in what form - it is subsequently removed, subjects that achieve greater efficacy may also be at risk of greater toxicity. This is particularly the case for drugs with a narrow therapeutic window - as described in Chapter One. Secondly from a purely practical standpoint, hospitals and surgeries find it difficult enough to prescribe or administer the correct drugs and doses for even straightforward regimens. For instance, over 6 months, 616 (5.7%) mistakes out of 10,778 written medication orders for children were detected at two US hospitals - the most common being a misplaced decimal place (Kaushal *et al*, 2001). Individualised treatment therefore represents an added level of complexity that current administrative systems would find seriously challenging.

Perhaps the initial utopian view is changing however, with sanity and realism beginning to return to drug development. According to Roses (2000), *The goal of pharmacogenomics is to account for and minimize interindividual variability in drug response, thus allowing clinicians to enhance the efficacy and minimize the toxicities associated with drug therapy.*

By considering the role of multiple genes, the field of pharmacogenomics seeks to divide a given population into smaller, less variable, more predictable subgroups, which enables clinicians to individualise drug therapy. Furthermore, from the regulatory perspective the expected potential for the incorporation of genetic information in drug development is expressed through the thoughts of Janet Woodcock from US's Center for Drug Evaluation and Research (CDER). She states that: *Through pharmacogenetics, hopefully, we will predict who will respond well to a drug. In addition, we want to be able to weed out people who will have serious side effects.* (Washington Drug Letter, 30 June 2003). The aim, therefore, is to establish links between specific genetic sequences and specific drug reactions that are, according to Woodcock, *dependable* and *fully researched.* These more pragmatic statements suggest therefore that rather than expecting new tailored-made treatments for individuals to be developed, the expectation is that only simple refinements to the labelling of current treatments is envisaged.

Of course statisticians have been using genetic information routinely in clinical research for many years through the investigation of the differential effects of gender. Furthermore, drug labelling has sometimes been adjusted to account for these observed differences. However it is true to say that there is indeed evidence of the increasing influence of genetics in drug development and labelling. For instance, the treatment of leukaemia with 6-mercaptopurine (6MP) can lead to severe myelosuppression at standard doses in the one in 300 people who have an extreme deficiency in the enzyme TPMT. The labelling will be amended as a result to state that a genetic test exists to identify those subjects who carry the gene on both chromosomes and that significant dose reduction and close monitoring is advised. For those heterozygous subjects (one in ten who carry the TPMT deficiency on just one chromosome) who produce less TPMT than normal, standard doses remain

acceptable (Pediatric Oncology Subcommittee of the FDA's Oncologic Drugs Advisory Committee (July 2003) as reported in Washington Drug Letter, 21 July 2003).

Like gender, race is an interesting area in relation to genetics since it is know that treatments sometimes have different effects in different races. For instance, two asthma drugs containing salmeterol required labelling changes to highlight an increased risk of life-threatening asthma attacks and death with these drugs in African-Americans (FDA as reported in The Wall Street Journal, 15 August 2003). However even though race is clearly of a genetic nature, confounding genes may actually prove to be a better predictor of outcome. For example, it has been found that the response to β-blockers varies with genotype and that the most responsive genotype is more widespread in Caucasians (42%) than in the Black population (18%). As such, although on average the Black population may have a poor response to β-blocker treatment, an individual Black person with the responsive genotype is likely to respond well. Indeed in the early 1990's, it was generally perceived that ACE inhibitors and β-blockers were ineffective in the Black population following results of the large Veterans Affairs Co-operative Study (Materson *et al*, 1993) - an observation that may now require refinement (JA Johnson speaking at the Annual Meeting of the American Society for Clinical Pharmacology and Therapeutics (2003) as reported in Clinical Psychiatry News, 1 July 2003). Similarly, in relation to the hepatic P-450 enzymes, around 75% of whites and 50% of blacks have a genetic inability to express functional CYP3A5 - part of the CYP3A family. Now, the effects of this difference are usually obscured since many drugs are metabolised by the universally expressed CYP3A4 - although this does lead to large between subject differences in the overall CYP3A activity. The clinical importance is as yet unclear but the potential for large differences in effect for specific drugs remain (Evans and McLeod, 2003).

Perhaps the goal of individualised treatment, afforded by developments in the area of genetics, should simply be viewed as a return to the values of the early 1900's as embodied by the Hippocratic ideal - that is, treatment of the *patient-as-a-person* (Porter, 2003). For instance Porter quotes Sir William Gull: *Never forget that it is not pneumonia, but a pneumonic man who is your patient.* (Ironically this movement was actually a reaction against the increasingly scientific approach being adopted by the Universities at the time and reflected a desire that healing should remain an art.) Reidenberg (2003) argues that current trends are simply a refinement of previous efforts by 18th Century physicians such as William Withering who tailored the use of digitalis to eliminate excess fluid. Indeed prior to 1900, few drug treatments were actually effective in anyone. Notable exceptions were digitalis itself (heart stimulation) together with quinine (malaria), opium (analgesic), colchicum (gout), amyl nitrate (angina) and, of course, aspirin (Porter, 2003). According to Reidenberg, *adjusting drug therapy to the individual has evolved from dose adjustments based on clinical effects to dose adjustments made in response to drug levels and now to dose adjustments based on deoxyribonucleic acid sequences of drug metabolizing enzyme genes.* This is indeed the pragmatist's view – with pharmacogenetics providing an opportunity for further refinement in our investigation of consistency of treatment effects.

In the final and concluding section of this Chapter and Research Thesis, focus will turn to the generalisability of the evidence that is accumulated from clinical trials and other sources. Some thoughts on the future for drug development are also presented.

7.5 GENERALISABILITY AND ROBUSTNESS

The true worth of the evidence generated from the clinical trial method is in its ability to make informed judgements regarding the treatment of future patients and it is in this context that the generalisation of results needs to be considered. According to ICH E9

(1998), generalisability is defined as *the extent to which findings of a clinical trial can be reliably extrapolated from the subjects who participated in the trial to a broader patient population and a broader range of clinical settings.* In this context, representation in clinical trials is an important aim and in relation to confirmatory trials, ICH E9 states that subjects should closely mirror the target population. According to Chatfield (2002), *Randomization is the means, but representativeness is the goal.* However as discussed in Chapter Two, simply ensuring a broad coverage of subjects is in itself insufficient to enable broad generalisation of results. According to Koch and Sollecito (1984), there are three areas to be considered when judging generalisability. Firstly, the coverage provided by the range of subjects encompassed by the target population; secondly, the demonstration that treatment differences are homogenous across investigators and the range of variation of demographic and pre-treatment characteristics; and thirdly, the replication of findings through multiple investigators.

Koch and Sollecito regard the basis for generalisability as strengthened when the coverage is as broad as possible in terms of geographic, demographic and pre-treatment characteristics. In terms of consistency of effect, they state: *When the treatment differences tend to be in the same direction and similar magnitude across the ranges of variation of such factors, the results of a study can be interpreted as having minimum dependence on the processes by which investigators and patients were selected to be included in it.* This brings in the concept of robustness which according to ICH E9, *refers to the sensitivity of the overall conclusions to various limitations of the data, assumptions, and analytic approaches to data analysis.* In other words, according to Koch and Sollecito, robustness supports generalisability. Finally, in relation to replication, support is provided for generalisability when similar findings are observed across multiple investigators. More broadly, the *reproducibility of valid conclusions in sufficiently many controlled settings*

reasonably supports generalization from judgementally defined study populations to a more extensive target population of conceptually similar patients.

Now, although both representation and consistency of effect have been discussed extensively in this Research Thesis, the concept of replication has not been raised previously. However before turning attention towards this concept it is informative to consider some other thoughts on the topic of generalisability. Davis (1994) discusses generalisability in wider terms and, in her view, data from seven different types of study are necessary to assess generalisability: laboratory (basic science); animal; genetic (if applicable); observational; clinical; epidemiological; and other RCTs with similar settings or treatments. One could argue that to this list should be added pharmacokinetic studies as, has been shown in Chapter Six, these are key to extrapolating data from adults to children. Lewis (1995) also argues that the *range and limits* of generalisability should be covered by medical and scientific considerations. That is, our understanding of the disease process and drug action *will usually permit satisfactory extrapolation of clinical trial results, provided the trial in question is soundly conducted and its results are convincing.* The view of Cowan and Wittes (1994) is that, *the closer an intervention is to a purely biological process, the more confident we feel in extrapolating beyond the types of patient studied.*

In relation to replication, there has been great debate regarding the need for more than one confirmatory study to support regulatory approval. The FDA's requirement for more than one study was based on the principle that one should be able to replicate findings. In was in addition to the FDA's established requirement for substantial evidence and the desire to be able to make generalisations to additional populations. (Note that some drugs have in fact been approved by the FDA with just one confirmatory study in cases were replication

was considered unethical.) If results were replicated – the argument went – then this demonstrated *future replicability* (Peck and Wechsler, 2002). Of course, replication can actually be achieved internally within a study through using multiple centres, countries and regions as described by Koch and Sollecito (1984) - it is simply a case of investigating consistency of effect. External validation, it is argued however, must come from other confirmatory studies - although Lewis (1995) has questioned whether this type of external replication is appropriate to pharmaceutical development. He argues that by the time phase III is reached *a confirmatory trial is just that - confirmatory of all the work which has gone before.* Chatfield (2002) considers replication as good scientific practice. That is, statisticians *need to give more emphasis to collecting more than one data set wherever possible, as that is the route to scientifically valid and generalizable results.* However, if generalisability is the real issue then the real requirement is for more varied evidence rather than more of the same. This view is supported by Howson and Urbach (1989) who state: *When an experiment's capacity to generate confirming evidence has been exhausted through repetition, further support would have to be sought from other experiments, moreover, experiments of different kinds.* And later: *Evidence that is varied is often regarded as offering better support to a hypothesis than an equally extensive volume of homogenous evidence.* Indeed Robert Temple of the FDA has acknowledged that the actual intention is that a second study would not be an *exact repetition* of the first (Peck and Wechsler, 2002). In fact the development of drugs in the paediatric population can be considered to fit well into this philosophy. That is, a study undertaken in children – by definition not an exact copy of what has gone before - has the potential not only to provide data for paediatric labelling but also adds to the weight of evidence to support the overall efficacy and safety of the drug in question. The real emphasis in drug development therefore should not be replication - which should be abandoned as a concept - but the accumulation of robust evidence in different subgroups, sub-populations and populations.

In this respect, drug development should be a systematic process of information gathering where studies (including PK investigations) begin with low risk groups and move to higher risk groups - paediatrics, the elderly, pregnant women, etc. Labelling then evolves through time as new and varied information is brought to the regulatory authorities. Of course such labelling could include dosing recommendations based on genetic information but the overall development task should not be deemed complete until all sections of society are covered. In this respect, labelling contraindications and exclusions should be based on actual scientific data rather than the absence of data. This concept is key to generating robust evidence and to the development of safe and effective treatments for society as whole.

Bailey (1994) states: *If all human beings were at the same risk and experience the same benefit from a given treatment, then we could generalise the results of a trial to any conceivable subset of people to the entire human population.* This, of course, would truly represent the Promised Land - a land of universal remedies. Indeed the idea of universal remedies has always been able to catch the public's imagination. Lydia E Pinkham's Vegetable Compound, which was sold from 1873, made Lydia Pinkham America's first millionairess. In fact "Lily the Pink", as she was known, was even celebrated in a song of the same name by the Liverpool band The Scaffold - *For she invented medicinal compound, Most efficacious in every case.* In the UK, Thomas Beecham manufactured his Pills and famous powders while James Morrison made a fortune from vegetable pills (Porter 2003) – although it is not known whether either of these was ever similarly feted by the rock and pop community.

Only time will tell whether the search for individualised drug treatment in the 21st Century proves to be as fruitless as the quest for universal remedies by Lily the Pink and others in the 19th Century. What is clear however, is that treatments that are broadly effective at a

single dose level are highly desirable in any society. It is likely therefore that current emphasis on generalisability of data across broad subject populations will continue in drug development and regulation - and it is hard to imagine this being satisfied without the accumulation of robust evidence covering all sections of society. In this respect thoughtful construction of analysis populations and subgroups will continue to play a key role for many years to come and it is hoped that this research thesis provides an insight as to how sub-setting might be used effectively such that the accumulation of robust evidence is achieved. Perhaps, therefore, the real challenge in the 21st Century is to develop safe and effective treatments that can be shown to be independent of genetic make-up and other external factors, and which can be administered both uniformly and simply - that is, most efficacious in every case. I'll drink a drink a drink to that!

SIMULATION NOTE

For pragmatic reasons, the simulations reported in chapters three and five were undertaken using 5000 trials. (Each individual simulation took in the region of 20 minutes to complete.) For this reason it is important to provide an indication of precision for the range of parameters estimated. As such exact 95% confidence intervals (Clopper-Pearson method) produced from StatXact (Cytel, 1999) together with standard errors (SE) calculated using the Normal approximation to the binomial distribution are given in the table below.

Percentage	Exact 95% CI	SE (Normal approximation)
0	0 to 0.07	Not defined
0.02	0 to 0.11	0.020
0.1	0.03 to 0.23	0.045
0.2	0.10 to 0.37	0.063
0.5	0.32 to 0.74	0.10
1	0.74 to 1.32	0.14
2.5	2.09 to 2.97	0.22
5	4.41 to 5.64	0.31
7.5	6.78 to 8.27	0.37
10	9.18 to 10.87	0.42
20	18.90 to 21.14	0.57
30	28.73 to 31.29	0.65
40	38.64 to 41.37	0.69
50	48.60 to 51.40	0.71
80	78.86 to 81.10	0.57

In the presentation of the simulation results, the data have been reported to two decimal places and the information above may be used to aid interpretation. Note that for the simulations of Simpson's paradox (SP) in Chapter Three, one aim was to identify any cases of SP under the conditions tested. As such, a reduction in the number of decimal places would have meant that some combinations would have been reported as having an incidence of zero.

REFERENCES

Alano MA, Ngougmna E, Ostrea EM, Konduri CG. Analysis of nonsteroidal anti-inflammatory drugs in meconium and its relation to persistent hypertension in the newborn. *Pediatrics* 2001; **107(3)**: 519-523.

Aldrich J. Correlations genuine and spurious in Pearson and Yule. *Statistical Science* 1995; **10**:364-376.

American Academy of Pediatrics: Committee on Drugs. Guidelines for the ethical conduct of studies to evaluate drugs in pediatric populations. *Pediatrics* 1995; **95(2)**:286-294.

Anello C. Emerging and recurrent issues in drug development. *Statistics in Medicine* 1999; **18**: 2301-2309.

AP Online: Press Association Inc. 26 August 2003.

The Arizona Republic: Gannett Co., Inc/ Newspaper Division. 5 September 2001.

Armitage P. Attitudes in clinical trials. *Statistics in Medicine* 1998; **17**:2675-2683.

Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *The Lancet* 2000; **355**:1064-1069.

Associated Press Newswires: Press Association Inc. 28 January 2001; 28 February 2001; 17 April 2001; 17 May 2001; 3 August 2001; 17 October 2001.

Association of the British Pharmaceutical Industry (2001). (www.abpi.org.uk/publications/publication_details).

Atz A, Wessel DL. Sildenafil ameliorates effects of inhaled nitric oxide withdrawal. *Anesthesiology* 1999; **91**:307-310.

Banner W. Off label prescribing in children: a view from the United States. *British Medical Journal* 2002; **324**:1290-1291.

Bauer P. Multiple testing in clinical trials. *Statistics in Medicine* 1991; **10**:871-890.

Bailey KR. Generalising the Results of Randomised Clinical Trials. *Controlled Clinical Trials* 1994; **15**:15-23.

Bennett JC. Special reports. Inclusion of women in clinical trials – policies for population subgroups. *New England Journal of Medicine* 1993; **329**: 288-292.

The Blue Sheet: F-D-C Reports Inc. 25 April 2001.

Blyth CR. On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association* 1972; **67**:364-366.

Bohidar NR, Peace KE. Pharmaceutical formulation development. In: Peace KE eds. *Biopharmaceutical statistics for drug development*. New York: Marcel Dekker, 1988:149-229.

Breslow NE. Statistics in the life and medical sciences. In: Raftery AE, Tanner MA, Wells MT, eds. *Statistics in the 21st Century*. Virginia: Chapman and Hall/CRC, 2001:1-3.

Bristol DR. Clinical equivalence. *Journal of Biopharmaceutical Statistics* 1999; **9**:549-561.

Brittain E, Lin D. A comparison of intent-to-treat and per-protocol results in antibiotic non-inferiority trials. *Statistics in Medicine* 2005; **24**: 1-10.

Report of a working party of the British Society of Antimicrobial Chemotherapy. The clinical evaluation of anti-bacterial drugs. *The Journal of Antimicrobial Chemotherapy* 1989; **23** Supplement B.

Bross I. Misclassification in 2x2 Tables. *Biometrics* 1954; **10**:478-486.

Brown DJ. ICH E9 guideline 'Statistical principles for clinical trials': a case study. Response to A Phillips and V Haudiquet. *Statistics in Medicine* 2003; **22**:13-17.

Budetti PP. Ensuring safe and effective medications for children. *Journal of the American Medical Association* 2003; **290**:950-951.

CHMP. Guideline on the clinical trials in small populations. (CHMP/EWP/83561/2005 draft) 17 March 2005.

CHMP. Notes for guidance on the clinical evaluation of vaccines. (CHMP/VWP/164653/2005) 17 May 2005.

Charig CR, Webb DR, Payne S.R, Wickham OE. Comparison of treatment of renal calculi by operative surgery, percutaneous nephrolithotomy, and extracorporeal shock wave lithotripsy. *British Medical Journal* 1986; **292**:879-882.

Chatfield C. Confessions of a pragmatic statistician. *The Statistician* 2002; **51 Part 1**:1-20.

Chicago Tribune: Chicago Tribune. 1 December 2000.

Choi SC, Lu IL. Effect of non-random missing data mechanisms in clinical trials. *Statistics in Medicine* 1995; **14**:2675-2684.

Chow S-C, Pong A. An overview of the regulatory approval process in drug development. *Drug Information Journal*, 1998; **32**:1175S-1185S.

Christie J. Case study - using western data in a Japanese submission. PSI Conference 2003.
http://www.psiweb.org/resources/show_resource_details.asp?id=1061&parentfolderid=393&itemtype=2&subgroup_id=3 [last accessed 10/07/03].

Clinical Psychiatry News: International Medical News Group. 1 July 2003.

Cohen J. *Statistical Power Analysis for the Behavioural Sciences*, revised edn. Academic Press, New York, 1977.

Committee on Children with Disabilities. American Academy of Pediatrics. The pediatrician's role in the diagnosis and management of autistic spectrum disorder in children. *Pediatrics* 2001; **107**:1221-1226.

Consensus panel of the Immunocompromised Host Society. The design, analysis and reporting of clinical trials on the empirical antibiotic management of the neutropenic patient. *The Journal of Infectious Diseases* 1990; **161**:397-401.

Coté CJ, Kauffman RE, Troendle GJ, Lambert GH. Budetti PP. Is the "Therapeutic Orphan" about to be adopted? *Pediatrics* 1995; **95**:118-123.

Cowan CD, Wittes J. Intercept studies, clinical trials and cluster experiments: to whom can we extrapolate. *Controlled Clinical Trials* 1994; **15**:24-29.

Cox D. *Analysis of Binary Data*. Chapman and Hall: London, 1970.

CPMP. Guideline on clinical investigation of hypnotic drugs. (III/3855/89) 1992.

CPMP note for guidance (III/3630/92-EN). Biostatistical methodology in clinical trials in applications for marketing authorizations for medicinal products. *Statistics in Medicine* 1995; **14**:1659-1682.

CPMP. Note for guidance on clinical investigation of medicinal products in children. (CPMP/EWP/462/95) 17 March 1997.
(http://www.emea.eu.int/pdfs/human/ewp/046295en.pdf) [last accessed 23/03/04].

CPMP. Note for guidance on evaluation of new anti-bacterial medicinal products. (CPMP/EWP/558/95) April 1997.
(http://www.emea.eu.int/pdfs/human/ewp/055895en.pdf) [last accessed 28/08/01].

CPMP. Concept paper on the development of a committee for proprietary medicinal products (CPMP) points to consider on biostatistical/methodological issues arising from recent CPMP discussions on licensing applications: choice of delta. (CPMP/EWP/2158/99) 23 September 1999.
(http://www.emea.eu.int/pdfs/human/ewp/215899en.pdf) [last accessed 28/08/01].

CPMP. Points to consider on biostatistical/methodological issues arising from recent CPMP discussions on licensing applications: superiority, non-inferiority and equivalence. (CPMP/EWP/482/99 draft) 23 September 1999.

CPMP. Points to consider on switching between superiority and non-inferiority. (CPMP/EWP/482/99) 27 July 2000.
(http://www.emea.eu.int/pdfs/human/ewp/048299en.pdf) [last accessed 28/08/01].

CPMP. Note for guidance on the investigation of bioavailability and bioequivalence. (CPMP/EWP/QWP/1401/98) 26 July 2001.
(http://www.emea.eu.int/pdfs/human/ewp/140198en.pdf) [last accessed 28/08/01].

CPMP. Concept paper on the development of a committee for proprietary medicinal products (CPMP) points to consider on the evaluation of the pharmacokinetics of medicinal products in the paediatric population. (CPMP/EWP/968/02) 30 May 2002.

CPMP. Note for Guidance on clinical investigation of medicinal products for the treatment of migraine. (CPMP/EWP/788/01) 19 September 2002. (http://www.emea.eu.int/pdfs/human/ewp/078801en.pdf) [last accessed 01/04/03].

CPMP. Concept paper on the pharmacovigilance for medicines used by children. (CPMP/PhVWP/4838/02) 17 October 2002.

CPMP. Points to consider on missing data. (CPMP/EWP/1776/99) 15 November 2001.

CPMP. Notes for guidance on the development of vaccinia virus based vaccines against smallpox. (CPMP/1100/02) 24 June 2002.

CPMP. Points to consider on multiplicity issues in clinical trials. (CPMP/EWP/908/99) 19 September 2002.

CPMP. Points to consider on adjustment for baseline covariates. (CPMP/EWP/2863/99) 22 May 2003.

CPMP. Note for guidance on evaluation of anticancer medicinal products in man: Addendum on paediatric oncology. (CPMP/EWP/569/02) 24 July 2003. (http://www.emea.eu.int/pdfs/human/ewp/056902en.pdf) [last accessed 23/03/04].

CPMP. Points to consider on the choice of non-inferiority margin. (CPMP/EWP/2158/99 draft) 26 February 2004.

CPMP. Guideline on the choice of the non-inferiority margin. (CPMP/EWP/2158/99 draft) 27 July 2005.

CYTEL Software Corporation, *StatXact 4 for Windows, User manual.* CYTEL Software Corporation: Cambridge, MA, July 1999.

Davis CE. Generalising from Clinical Trials. *Controlled Clinical Trials* 1994; **15**:11-14.

Day S. Changing times in pharmaceutical statistics: 2000-2020. *Pharmaceutical Statistics* 2002; **1**:75-82.

Device and Diagnostic Letter: Washington Business Information Inc. 30; No. 29, 28 July 2003.

Doull IJM. Review: The effect of asthma and its treatment on growth. *Archives of Diseases in Childhood* 2004; **89**:60-63.

Dunnett CW, Gent M. Significance testing to establish equivalence between treatments, with special reference to data in the form of 2x2 tables. *Biometrics* 1977; **33**:593-602.

Ebbutt AF, Frith L. Practical issues in equivalence trials. *Statistics in Medicine* 1998; **17**:1691-1701.

EFSPI Working Group. Qualified statisticians in the European pharmacaeutical industry: Report of a European Federation of Statisticians in the Pharmaceutical Industry (EFSPI) Working Group. *Drug Information Journal*, 1999;**33**:407-415.

Ellenberg JH. Intent-to-treat analysis versus as treated analysis (with discussion). *Drug Information Journal* 1996; **30**:535-544.

Elston RC, Idury RM, Cardon LR, Lichter JB. The study of candidate genes in drug trials: sample size considerations. *Statistics in Medicine* 1999; **18**:741-751.

EMEA. Overview of the European authorisation system (2002). (http://www.eudra.org/aboutus.htm) [last accessed 14/09/02].

Evans WE, McLeod HL. Pharmacogenomics - drug disposition, drug targets and, side effects. *The New England Journal of Medicine* 2003; **348**:538-549.

Farrington CP, Miller E. Review: Vaccine trials. *Molecular Biotechnology* 2001; **17**:43-58.

FDA. Guideline for the format and content of the clinical and statistical sections of an application, July 1988. (http://www.fda.gov/cder/regulatory/applications/guidance.htm) [last accessed 11/02/02].

FDA. Guidance for industry. Evaluating clinical studies of antimicrobials in the division of anti-infective drug products. 17 February 1997. (http://www.fda.gov/cder/guidance/draft9al.pdf) [last accessed 24/08/01].

FDA. Guidance for industry. Bioavailability and bioequivalence studies for orally administered drug products - general considerations. October 2000. (http://www.fda.gov/cder/guidance/3615fnl.htm) [last accessed 15/05/02].

FDA. Guidance for industry. Statistical approaches to establishing bioequivalence. January 2001. (http://www.fda.gov/cder/guidance/3616fnl.htm) [last accessed 15/05/02].

FDA CDER (2002). Preventable adverse drug reactions: a focus on drug interactions. (http:/www.fda.gov/cder/drug/drugReactions/default.htm) [last accessed 20/09/02].

FDA. FDA History (2002). (http://www.fda.gov/oc/history/default.htm) [last accessed 29/10/02].

FDA. FDA Talk Paper. FDA Discusses pediatric drug safety with its advisory subcommittee under Best Pharmaceuticals for Children Act. 2003. (http://63.240.199.20/bbs/topics/ANSWERS/2003/ANS01228.html).

Federal Register / Vol 60, No. 174 / Friday, September 8, 1995 / Proposed Rules 46794-46796.

Feinstein AR. Intent-to-treat policy for analyzing randomized trials: Statistical distortions and neglected clinical challenges. In *Patient Compliance in Medical Practice and Clinical Trials*, Cramer JA and Spilker B (ed). Raven Press, Ltd: New YorkYork, 1991; 359-370.

Fisher LD, Dixon DO, Herson J et al. Intention to treat in clinical trials. In *Statistical issues in drug research and development*, Peace KE (ed). Marcel Dekker: New York, 1990; 331-350.

Fleiss JL. *The design and analysis of clinical experiments*. John Wiley & Sons, New York, 1986.

Ford I, Norrie J, Atimadis S. Model inconsistency, illustrated by the Cox proportional

Francis T, Korns RF, Voight T, Boisen M, Hemphill FM, Napier JA, Tolchinsky E. An evaluation of the 1954 poliomyelitis vaccine trials. *Am J Public Health* Part 2 1955; **45**:1-63.

Freeman JV, Cole TJ, Chinn S, Jones PRM, White EM, Preece MA. Crosssectional stature and weight reference curves for the UK. *Arch Dis Child* 1995; **73**:17-24.

Frith L. Experiences of drug development in the Japan environment. PSI Conference 2003.
http://www.psiweb.org/resources/show_resource_details.asp?id=1060&parentfolderid=393&itemtype=2&subgroup_id=3 [last accessed 10/07/03].

Gail MH. Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. In: *Modern Statistical Methods in Chronic Disease Epidimiology*, Moolgavkar SH, Prentice RL (eds). Wiley: New York, 1986; 3-18.

Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 1985; **41:** 361-372.

Gambian Hepatitis Study Group. The Gambia hepatitis intervention study. *Cancer Research* 1987; **47**: 5782-5787.

Garrett AD. Therapeutic equivalence: fallacies and falsification. *Statistics in Medicine* 2003; **22**: 741-762.

Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 1990; **85:** 398-409.

Gelman A, Rubin DB. Markov chain Monte Carlo methods on biostatistics. *Statistical Methods in Medical Research* 1996; **5:** 339-355.

Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1984; **6:** 721-741.

Generic line: Washington Business Information Inc. 7 September 2001.

Gillings D, Koch G. The application of the principle of intention-to-treat to the analysis of clinical trials. *Drug Information Journal* 1991; **25**:411-424.

Goldberg J. The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. *Journal of the American Statistical Association* 1975; **70:** 561-567.

Hand DJ. Deconstructing statistical questions (with discussion). *Journal of the Royal Statistical Society. Series A* 1994; **157:**317-356.

Harrington DP. The randomized clinical trial. In *Statistics in the 21st Century*, Raftery AE, Tanner MA, Wells MT (eds). Chapman and Hall/CRC: Virginia, 2001; 67-74.

Health News Daily: F-D-C Reports Inc. 8 December 2000; 21 June 2001.

Hempel CG. Philosophy of Natural Science. Englewood Cliffs. NJ:Prentice Hall, 1966.

Hill A Bradford. *Principles of Medical Statistics.* 7th edn. London: The Lancet, 1961; 259.

Hjalmarson A, Herlitz J, Malek I *et al.* Effect on mortality of metoprolol in acute myocardial infarction. *The Lancet* 1981; 823-827.

Hollis S, Campbell F. What is meant by intention-to-treat analysis? Survey of published randomised controlled trials. *British Medical Journal* 1999; **319**:670-674.

Holm S. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* 1979; **6**:65-70.

Howson C, Urbach P. *Scientific Reasoning: A Bayesian Approach*. Open Court, La Salle: Illinois. 1989.

Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine* 1998; **17**:1623-1634.

Hughes WT, Pizzo PA, Wade JC, Armstrong D, Webb CD, Young LS. General guidelines for the evaluation of new anti-infective drugs for the treatment of febrile episodes in neutropenic patients: Evaluation of new anti-infective drugs for the treatment of febrile episodes in neutropenic patients. *Clinical Infectious Diseases* 1992; **15(1):**S206-S215.

Hutton JL. Number needed to treat: properties and problems (with comments). *Journal of the Royal Statistical Society, Series* A, 2000; **163:**403-419.

ICH topic E3. Note for guidance on structure and content of clinical study reports (CPMP/ICH/137/95). December 1995. (http://www.emea.eu.int/pdfs/human/ich/013795en.pdf) [last accessed 02/06/03].

ICH topic E5. Note for guidance on ethnic factors in the acceptability of foreign clinical data. (CPMP/ICH/289/95). 18 March 1998. (http://www.emea.eu.int/pdfs/human/ich/028995en.pdf) [last accessed 06/01/04].

ICH topic E6. Guideline for good clinical practice (CPMP/ICH/135/95). July 1996. (http://www.emea.eu.int/pdfs/human/ich/013595en.pdf) [last accessed 18/12/04].

ICH topic E7. Note for guidance on studies in support of special populations: geriatrics. (CPMP/ICH/379/95). March 1994. (http://www.emea.eu.int/pdfs/human/ich/037995en.pdf) [last accessed 06/01/04].

International Conference on Harmonisation. Statistical principles for clinical trials (ICH E9). *Statistics in Medicine* 1999; **18**:1905-1942.

ICH topic E10. Choice of control group in clinical trials (CPMP/ICH/364/96). 27 July 2000. (http://www.emea.eu.int/pdfs/human/ich/036496en.pdf) [last accessed 28/08/01].

ICH topic E11. Note for guidance on clinical investigation of medicinal products in the paediatric population. (CPMP/ICH/2711/99). 27 July 2000. (http://www.emea.eu.int/pdfs/human/ich/271199en.pdf) [last accessed 06/01/04].

ICH Topic M4. Common technical document for the registration of pharmaceuticals for human use - organisation of the common technical document (CPMP/ICH/2887/99). 26 November 2000. (http://www.emea.eu.int/pdfs/human/ich/288799enm.pdf) [last accessed 29/11/01].

Institute of Medicine Report. Childhood cancer survivorship: Improving care and quality of life. 26 August 2003.

Johnson JD. Past and present regulatory aspects of drug development. In: Peace KE, eds. *Biopharmaceutical statistics for drug development*. New York: Marcel Dekker, 1988:1-19.

Johnson-Pratt LR, Bush J. Activities of the pharmaceutical industry relative to the FDA gender guideline. *Drug Information Journal* 1996; **30**:709-714.

Julious SA, Mullee MA. Confounding and Simpson's paradox. *British Medical Journal* 1994; **309**:1480-1481.

Kaushal R, Bates DW, Landrigan C, McKenna KJ, Clapp MD, Federico F, Goldmann DA. Medication errors and adverse drug events in pediatric inpatients. *JAMA* 2001; **285**:2114-2120.

Kearns GL. Introduction: Drug development for infants and children: Rescuing the therapeutic orphan. *Drug Information Journal* 1996; **30**:1121-1123.

Kelly PJ, Stallard N, Whittaker JC. Statistical design and analysis of pharmacogenetic trials. *Statistics in Medicine* 2005; **24**:1495-1508.

Kirkwood BR, Cousens SN, Victora CG, de Zoysa I. Issues in the design and interpretation of studies to evaluate the impact of community-based interventions. *Tropical Medicine and International Health* 1997; **2**:1022-1029.

Koch G. In: Ellenberg S. Methodological issues in pivotal trials: discussion. *Drug Information Journal* 1996; **30**:563-565.

Koch G. Discussion of "p-value adjustments for subgroup analyses". *Journal of Biopharmaceutical Statistics* 1997; 7**(2)**:323-331.

Koch G, Davis SM, Anderson RL. Methodological advances and plans for improving regulatory success for confirmatory studies. *Statistics in Medicine* 1998; **17**:1675-1690.

Koch G, Gansky SA. Statistical considerations for multiplicity in confirmatory protocols. *Drug Information Journal* 1996; **30(2)**:523-534.

Koch G, Sollecito WA. Statistical considerations in the design, analysis, and interpretation of comparative clinical trials. *Drug Information Journal* 1984; **18**:131-151.

Lachin JM. Statistical properties of randomisation in clinical trials. *Controlled Clinical Trials* 1998; **9**:289-311.

Lane PW, Nelder JA. Analysis of covariance and standardization as instances of prediction. *Biometrics* 1982; **38**:613-621.

Lee PM. *Bayesian statistics: An introduction.* Edward Arnold: London, 1989.

Lee PN. Simple methods for checking for possible errors in reported odds ratios, relative risks and confidence intervals. *Statistics in Medicine* 1999; **18**:1973-1981.

Leeder J S. Developmental aspects of drug metabolism in children. *Drug Information Journal* 1996; **30**:1135-1143.

Lewis JA. Statistical issues in the regulation of medicines. *Statistics in Medicine* 1995; **14**:127-136.

Lewis JA, Jones DR, Röhmel J. Biostatistical methodology in clinical trials – a European guideline. *Statistics in Medicine* 1995; **14**:1655-1682.

Lewis JA. Editorial: Statistics and statisticians in the regulation of medicines. *Journal of the Royal Statistical Society, Series A* 1996; **159**:359-365.

Lewis JA, Facey KM. Statistical shortcomings in licensing applications. *Statistics in Medicine* 1998; **17**:1663-1673.

Lewis JA, Machin D. Intention to treat – who should use ITT? *British Journal of Cancer* 1993; **68**:647-650.

Lewis JA, Louv W, Rockfold F, Sato T. The impact of the international guideline entitled Statistical Principles for Clinical Trials (ICH E9). *Statistics in Medicine* 2001; **20**:2549-2560.

Lindley DV, Smith AFM. Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society. Series B* 1972; **34**:1-41.

Lindley DV, Novick MR, The role of exchangeability in inference. *The Annals of Statistics* 1981; **9**:45-58.

Little RJA, Rubin DB. *Statistical Analysis with Missing Data,* Wiley: New York, 1987.

Machin D, Campbell MJ. *Statistical tables for the design of clinical trials.* Blackwell Scientific Publications: Oxford, 1987.

Makuch RW, Simon R. Sample size requirements for evaluating a conservative therapy. *Cancer Treat Reports* 1978; **62**:1037-1040.

Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655-660.

Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal National Cancer Institute* 1959; **22**:719-748.

Marty M, Gershanovich M, Campos B et al. Letrozole, a new potent selective aromatase inhibitor (AI) superior to aminoglutethimide (AG) in postmenopausal women with advanced breast cancer (ABC) previously treated with anti-estrogens. *American Society of Clinical Oncology. Proceedings of 33rd Annual Meeting* 1997.

Matcham J. The design and analysis of a pivotal non-inferiority study - a case study. PSI Conference 2003. http://www.psiweb.org/resources/show_resource_details.asp?id=1042&parentfolderid=393&itemtype=2&subgroup_id=3 [last accessed 10/07/03].

Materson BJ, Reda DJ, Cushman WC et al, for The Department of Veterans Affairs Cooperative Study Group of Antihypertensive Agents. Single-drug therapy for hypertension in men: A comparison of six antihypertensive agents with placebo. *NEJM* 1993; **328**:914-921.

Meinert CL, Gilpin AK. Estimation of gender bias in clinical trials. *Statistics in Medicine* 2001; **20**:1153-1164.

Meldrum M. "A calculated risk": the Salk polio vaccine field trials of 1954. *British Medical Journal* 1998; **317**:1233-1236.

Merkatz RB. Special reports. Women in clinical trials of new drugs: a change in Food and Drug Administration policy. *New England Journal of Medicine* 1993; **329**: 292-296.

McRorie T. Quality drug therapy in children: Formulations and delivery. *Drug Information Journal* 1996; **30**:1173-1177.

MHRA. (www.medicines.mhra.gov.uk/ourwork/licensingmeds/types/clintrialdir.htm) [last accessed 22/12/04].

Milne C-P. Guest editor's note: The pediatric studies initiative: Solution to worldwide need? *Drug Information Journal* 2000a; **34**:193-195.

Milne C-P. The health of the world's children: What goes around comes around. *Drug Information Journal* 2000b; **34**:213-221.

Morikawa T, Yoshida M. A useful testing strategy in phase III trials: combined test of superiority and test of equivalence. *Journal of Biopharmaceutical Statistics* 1995; **5**:297-306.

National Center for Health Statistics 2003. (http://www.cdc.gov.nchs/about/major/nhanes/growthcharts/background.htm)

Nelder JA. The statistics of linear models: back to basics. *Statistics and Computing* 1994a; **4**:221-234.

Nelder JA. In: Hand DJ. Deconstructing statistical questions (with discussion). *Journal of the Royal Statistical Society. Series A*, 1994b; **157**:317-356.

Nelder JA, Wedderburn RWM. Generalised linear models. *Journal of the Royal Statistical Society, Series* A, 1972; **135**:370-384.

The New Shorter Oxford English Dictionary, Volume 2. Clarendon Press: Oxford, 1993.

The New York Times: The New York Times Digital. 11 February 2001.

National Institutes of Health Revitalization Act of 1993. §131, Pub L No 103-43, 107 Stat 133 (codified at 42 USC §289a-2).

Neglia JP, Friedman DL, Yasui Y, Mertens AC, Hammond S, Stovall M, Donaldson SS, Meadows AT, Robison LL. *Journal of the National Cancer Institute* 2001; **93**:618-629.

Newell DJ. Intention-to-treat analysis: Implications for quantitative and qualitative research. *International Journal of Epidemiology* 1992; **21**:837-841.

Norwood P. Clinical trials in biotechnology: A perspective form the pharmaceutical industry. *Drug Information Journal* 1996; **30**:559-5562.

Pan G, Wolfe DA. Test for qualitative interaction of clinical significance. *Statistics in Medicine* 1997; **16**:1645-1652.

Patja A, Davidkin I, Kurki T, Kallo MJ, Valle M, Peltola H. Serious adverse events after measles-mumps-rubella vaccination during a fourteen-year prospective follow-up. *Pediatr. Infect. Dis J* 2000; **19**:1127-1134.

Pearl J. *Causality: models, reasoning, and inference*. Cambridge University Press: Cambridge, 2000.

Pearson K, Lee A, Bramley-Moore L. Genetic (reproductive) selection: inheritance of fertility in man. *Philosophical Transactions of the Royal Society Series A* 1899; **192**:257-330.

Peck CC, Wechsler J. Report of a Workshop on confirmatory evidence to support a single clinical trial as a basis for new drug approval. *Drug Information Journal* 2002; **36**:517-534.

Peto R. Statistical aspects of cancer trials. In: Halman KE, eds. *Treatment of cancer.* London: Chapman and Hall: London, 1982; 867-871.

Peto R. Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomised clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer* 1976; **34**:585-612.

Peto R. Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomised clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *British Journal of Cancer* 1977; **35**:1-39.

Pharmaceutical Approvals Monthly: F-D-C Reports Inc. 1 January 2001; 1 May 2001.

Pharmaceutical Executive: Advanstar Communications Inc. 1 February 2000.

Phillips A, Ebbutt A, France L, Morgan D. The international conference on harmonization guideline "statistical principles for clinical trials": issues in applying the guideline in practice. *Drug Information Journal* 2000; **34:** 337-348.

Phillips A, Haudiquet V. ICH E9 guideline 'Statistical principles for clinical trials': a case study. *Statistics in Medicine* 2003; **22**:1-11.

Piantadosi S, Gail MH. A comparison of the power of two tests for qualitative interactions. *Statistics in Medicine* 1993; **12**:1239-1248.

The Pink Sheet: F-D-C Reports Inc. 30 July 2001; 22 October 2001.

Pocock SJ. *Clinical trials: a practical approach.* John Wiley & Sons: Chichester, 1983.

Pocock SJ, Altman S, Armitage P, Ashby D, Bland M, Chilvers C, Dawid P, Ebbutt A, Evans S, Finney D, Gardner M, Gore S, Jones D, Lewis J, Machin D, Matthews J, Spiegelhalter D, Sutherland I, Thompson S. Statistics and statisticians in drug regukation in the United Kingdom. *Journal of the Royal Statistical Society, Series* A 1991; **154**:413-419.

Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine* 2002; **21**:2917-2930.

Pong A, Chow S-C. Statistical/practical issues in clinical trials. *Drug Information Journal*, 1997; **31**:1167-1174.

Porter R. Blood and guts: A short history of medicine. Penguin Books: London, 2003.

PR Newswire: PR Newswire Association Inc. 10 March 2000; 10 May 2001; 13 June 2001; 23 June 2001; 7 April 2003.

Racine A, Grieve AP, Flühler H, Smith AFM. Bayesian methods in practice: experiences in the pharmaceutical industry (with Discussion). *Applied Statistics* 1986; **35**:93-150.

Reed MD. The ontogeny of drug disposition: focus on drug absorption, distribution, and excretion. *Drug Information Journal* 1996; **30:**1129-1134.

Reidenberg MM. Commentaries: Evolving ways that drug therapy is individualized. *Clinical Pharmacology and Therapeutics* 2003; **74**:197-202.

Roberts R. Adequacy of current laws governing research in children: the view from FDA. 2002. (http://www.fda.gov/cder/pediatric/presentation/SFEthics/tsld001.htm) [last accessed 28/05/02].

Roberts R, Maldonado S. FDA center for drug evalaution and research (CDER) pediatric plan and new regulations. *Drug Information Journal* 1996; **30:**1125-1127.

Roberts R, Rodriguez W, Murphy D, Crescenzi T. Pediatric drug labelling: Improving the safety and efficacy of pediatric therapies. *Journal of the American Medical Association* 2003; **290:**905-911.

Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review* 1991; **58**(2):227-240.

Robinson LD, Dorroh JR, Lein D, Tiku ML. The effects of covariate adjustment in generalized linear models. *Communications in Statistics. – Theory and Methods* 1998; **27**:1653-1675.

Röhmel J. Therapeutic equivalence investigations: statistical considerations. *Statistics in Medicine* 1998; **17**:1703-1714.

Röhmel J. Controversies about sponsor initiated re-analysis of clinical trial data in the licensing process. *Statistics in Medicine* 1999; **18**:2321-2330.

Röhmel J. Statistical considerations of FDA and CPMP rules for the investigation of new anti-bacterial products. *Statistics in Medicine* 2001; **20**:2561-2571.

Roses AD. Pharmacogenetics and the practice of medicine. *Nature* 2000; 405: 857-865.

Rothmann M, Li N, Chen G, Chi YH, Temple R, Tsou H-H. Design and analysis of non-inferiority mortality trials in oncology. *Statistics in Medicine* 2003; **22**:239-264.

Royal College of Paediatrics and Child Health. Expert Consensus Group. 18 February 2002 (www.rcpch.ac.uk/publications/recent_publications/finalgrowth.pdf).

SAS Institute Inc., *SAS/STAT® User's Guide, Version 6, fourth, Volume 2,* Cary, NC: SAS Institute Inc., 1989.

SAS Institute Inc. *SAS/STAT® Software: Changes and enhancements through release 6.11.* SAS Institute Inc.: Cary, NC, 1996.

Schaller J G. Drugs for children: The world situation. *Drug Information Journal* 2000; **34**:197-201.

Schultz JR, Ruppel PL, Johnson MA. Pharmaceutical lead discovery and optimization. In *Biopharmaceutical statistics for drug development*, Peace KE (ed). Marcel Dekker: New York, 1988; 21-82.

Schwartz D, Flamant R, Lellouch J. *Clinical trials.* Academic Press: London, 1980.

Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in clinical trials. *Journal of Chronic Disease* 1967; **20**:637-48.

Selwyn MR. Preclinical safety development. In: Peace KE, eds. *Biopharmaceutical statistics for drug development.* Marcel Dekker: New York, 1988:231-271.

Senn S. Falsificationism and clinical trials. *Statistics in Medicine* 1991; **10**:1679-1692.

Senn S. *Cross-over trials in clinical research.* John Wiley & Sons: Chichester, 1993.

Senn S. Testing for baseline balance in clinical trials. *Statistics in Medicine* 1994a; **13**:1715-1726.

Senn S. Fisher's game with the devil. *Statistics in Medicine* 1994b; **13**:217-230.

Senn S. *Statistical issues in drug development,* John Wiley and Sons: New York, 1997.

Senn S. Letter to the Editor. *SPIN* September 1999: 4-5.

Senn S. Consensus and controversy in pharmaceutical statistics (with discussion). *Journal of the Royal Statistical Society, Series D* 2000; **49**: 135-176.

Senn S. Individual therapy: new dawn or false dawn? *Drug Information Journal* 2001a; **35**:1479-1494.

Senn S. Statistical issues in bioequivalence. *Statistics in Medicine* 2001b; **20**:2785-2799.

Senn S. Guest Editorial: The misunderstood placebo. *Applied Clinical Trials* 2001c; **10**(5):40-46.

Senn S. Ethical considerations concerning treatment allocation in drug development trials. *Statistical Methods in Medical Research* 2002; **11**:403-411.

Senn S. *Dicing with death: chance, risk and health.* Cambridge University Press: Cambridge, 2003.

Simpson EH. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B* 1951; **13**:238-241.

Shirkey H. Editorial comment: therapeutic orphans. *J Pediatr.* 1968; **72**:119-120.

Silverman WA, Chalmers I. Sir Austin Bradford Hill: An Appreciation. *Controlled Clinical Trials* 1992; **13**:100-105.

Simon R, Freedman LS. Bayesian design and analysis of two x two factorial clinical trials. *Biometrics* 1997; **53**:456-464.

Simon R. Bayesian subset analysis: application to studying treatment-by-gender interactions. *Statistics in Medicine* 2002; **21**:2909-2916.

Sleight P. Commentary. Debate: Subgroup analyses in clinical trials - fun to look at, but don't believe them! *Curr Control Trials Cardiovasc Med* 2000; **1**:25-27.

Smith AFM, Roberts GO. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 1993; **55**:3-24.

Smith C, Burley C, Ireson M. Clinical trials of antibacterial agents: a practical guide to design and analysis. *Journal of Antimicrobial Chemotherapy* 1998; **41**:467-480.

Smith P, Kerr GD, Cockel R, Ross BA. et al. A comparison of omeprazole and ranitidine in the prevention of recurrence of benign osophageal stricture. *Gastroenterology* 1994; **107**:1312-1318.

Smith PG, Hayes RJ. Design and conduct of field trials of malaria vaccines. In: Targett GAT eds. *Malaria: Waiting for the Vaccine.* Wiley & Sons, 1991:199-215.

Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomised trials (with discussion). *Journal of the Royal Statistical Society, Series* A 1994; **157**:357-416.

Spiegelhalter DJ, Thomas A, Best NG, Lunn D. WinBUGS, Version 1.4, User Manual. MRC Biostatistics Unit: Cambridge, 2001.

Spielberg SP. Opportunities for pediatric drug development: the knowledge bridge from basic science to clinical application. *Drug Information Journal* 1996; **30**:1145-1148.

Swarbrick ET, Gough AL, Foster CS, Christian J, Garrett AD and Langworthy CH. Prevention of recurrence of oesophageal stricture, a comparison of lansoprazole and high-dose ranitidine. *European Journal of Gastroenterology and Hepatology* 1996; **8**:431-438.

Tan Sheet: F-D-C Reports Inc. 3 September 2001.

The Times-Picayune: Times-Picayune Publishing Corp. 28 March 2001.

Testimony: Pediatric drug research, (GAO-01-705T). United States General Accounting Office. May 8th 2001.

Trials of war criminals before the Nuremberg Military Tribunals under Control Council Law. Nuremberg, October 1946 – April 1949. Washington, DC: US Government Printing Office, 1949; **10 (2)**:181-182.

Tu D. On the use of the ratio or the odds ratio of cure rates in therapeutic equivalence clinical trials with binary endpoints. *Journal of Biopharmaceutical Statistics* 1998; **8**:263-282.

Tufts CDD. Tufts Center for the Study of Drug Development pegs costs of a new prescription medicine at $802 million. Press release. 30 November 2001. (http://csdd.tufts.edu/NewsEvents/RecentNews.asp?newsid=6) [last accessed 24/03/03].

United States General Accounting Office. Testimony: Pediatric drug research. Substantial increase in studies of drugs for children, but some challenges remain. GAO-01-750T. 8 May 2001

Unnebrink K, Windeler J. Sensitivity analysis by worst and best case assessment: is it really sensitive? *Drug Information Journal* 1999; **33**:835-839.

USA Today: USA Today Information Network. 20 December 2000.

The Wall Street Journal: Dow Jones & Company Inc. 5 February 2001.

Wellek S. Testing for absence of qualitative interactions between risk factors and treatment effects. *Biometrical Journal* 1997; **39**:809-821.

Washington Drug Letter: Washington Business Information Inc. 21 July 2003; 28 July 2003; 15 August 2003.

Whitehead J. Sample size calculations for ordered categorical data. *Statistics in Medicine* 1993; **12**:2257-2271.

World Health Organisation. Global Advisory Committee on Vaccine Safety. Statement on thiomersal. August 2003. (http://www.who.int/vaccine_safety/topics/thiomersal/statement200308/en/index.html) [last accessed 07/12/05].

World Health Organisation Handbook for Reporting Results of Cancer Treatment. World Health Organisation, Geneva, 1979. WHO Offset Publication No. 48.

Wilson JT. Strategies for pediatric drug evaluation: a view from the trenches. *Drug Information Journal* 1996; **30**:1149-1162.

World Medical Association, 52nd Assembly, Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects (World Medical Association, Inc., Edinburgh, 2000.(www.wma.net).

Yan X. Test for qualitative interaction in equivalence trials when the number of centres is large. *Statistics in Medicine* 2004; **23**:711-722.

Zidek J. Maximal Simpson-disaggregations of 2x2 tables. *Biometrika* 1984; **71**:187-190.